FINAL REPORT

The UXO Classification Demonstration at Former Camp Butner, NC

ESTCP Project

July 2011

Shelley Cazares
Michael Tuley
Elizabeth Ayers
Institute for Defense Analysis

ESTCP

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE
JUL 2011

2. REPORT TYPE
N/A

3. DATES COVERED
-

4. TITLE AND SUBTITLE
The UXO Classification Demonstration at the Former Camp Butner, NC

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Defense Analysis

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
ESTCP Project, The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:

a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
unclassified	**unclassified**	**unclassified**	**SAR**	**164**	

About This Publication

This work was conducted by the Institute for Defense Analyses (IDA) under contract DASW01-04-C-0003, Task AM-2-1528, "Surface and Buried UXO Detection," for the Strategic Environmental Research and Development Program (SERDP) and Environmental Security Technology Certification Program (ESTCP). The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Copyright Notice

4850 Mark Center Drive, Alexandria, Virginia 22311-1882 • (703) 845-2000.

INSTITUTE FOR DEFENSE ANALYSES

IDA Document D-4379

The UXO Classification Demonstration at the Former Camp Butner, NC

Shelley Cazares
Michael Tuley
Elizabeth Ayers

Executive Summary

Introduction

Unexploded ordnance (UXO) are munitions that were armed and fired but did not explode. Their risk of detonation remains, even decades after initial use. Thousands of sites in the United States are suspected of UXO contamination and require remediation. As much as 75% of current remediation costs may be associated with digging up nonhazardous scrap metal called “clutter,” instead of UXO. The development, validation, and acceptance of reliable technologies to correctly classify buried targets as UXO or clutter could lead to a significant reduction in UXO remediation costs, allowing more land to be cleared for the same amount of funding.

The Environmental Security Technology Certification Program (ESTCP) carried out the third live-site UXO classification demonstration at the former Camp Butner, NC in 2010. The main goal of the demonstration was to test and validate currently available and emerging classification technologies on a live site under operational conditions. Another goal was to involve environmental regulators, program managers, and other stakeholders in the design, execution, and evaluation of the demonstration to better understand what might be required in a real-world remediation project if detected targets were classified as clutter and therefore left in the ground.

Methods

The former Camp Butner was chosen to provide a greater challenge than the previous two demonstration sites. Historical records showed that a variety of munitions had been fired at Butner, including 37 mm projectiles similar in size to clutter. Thus, the estimated size of the buried target was not likely to be a discriminating feature in this demonstration, unlike in the previous two.

Although this was a live-site demonstration, 160 inert UXO were seeded at the site, as in the previous two demonstrations. Clutter is common at live sites, but UXO is rare. Additional UXO must be seeded to confidently assess classification capabilities against a well-characterized and statistically significant set of targets.

Three electromagnetic induction (EMI) instruments were used to collect data. The traditional EM61-Mk2 cart consists of one electronic coil that transmits an electromagnetic field and another coil that receives a secondary field. The more advanced MetalMapper consists of three orthogonal transmit coils as well as seven tri-axial receive coils. Similarly, the Time-domain Electromagnetic Multi-Sensor Towed Array Detection System (TEMTADS) consists of a 5 × 5 array of transmit and receive coils. Both the traditional EM61-Mk2 cart and the more advanced MetalMapper collected data in dynamic mode. Anomalies were selected in these data based on geophysical models of the smallest expected UXO buried at its deepest expected depth and most unfavorable orientation. The MetalMapper and TEMTADS then collected high-resolution static data for every detected anomaly.

Ground truth was carefully compiled. A commercial UXO remediation company recovered all targets from all anomalies. The ground truth label of "TOI" for target of interest was assigned to all anomalies from which at least one seeded or native UXO was recovered. All other anomalies (those with only clutter) were labeled "Non-TOI."

Six teams performed classification analyses as part of the demonstration. These included commercial geophysics companies (CH2M HILL and NAEVA Geophysics Inc.), as well as organizations involved in the research and development of advanced UXO classification technologies (Dartmouth College, Geometrics Inc., SAIC, and Sky Research Inc.). Two other companies (Parsons and Signals Innovations Group) performed retrospective analyses after ground truth had been released to all participants.

Different teams classified the detected anomalies using different methods applied to different data sets. Several different software suites were used to analyze the data, including UXOLab and the UX-Analyze module of Oasis montaj. Each classification analysis resulted in one ranked anomaly list. Anomalies were listed in rank order, from most to least likely Non-TOI. A "don't dig threshold" was then selected. In a real remediation project, all targets below threshold would be recovered, or "dug," while all targets surpassing threshold would not have to be dug and could remain in the ground.

IDA scored 54 ranked anomaly lists by comparing them to full ground truth. A curve was created for each list, similar to the receiver-operating characteristic (ROC) curves used in general classification problems. The curves and resulting statistics led to the findings and recommendations of this demonstration.

Key Findings

- **The EM61-Mk2 cart showed better detection performance than the MetalMapper in dynamic mode.** The EM61-Mk2 cart detected all seeded UXO, resulting in a probability of detection (Pd) of 100%. In comparison, the MetalMapper failed to detect two 37 mm projectiles seeded at 30 cm, resulting in a Pd of 99%. Although the differences in Pd were not statistically significant, had this been a real remediation project, stakeholders would have been troubled by the MetalMapper's failure to detect two seeds.

- **The EM61-Mk2 cart showed poor classification performance.** The EM61-Mk2 analyses led to the incorrect classification of many TOIs and/or small reductions in Non-TOI digs.
- **The MetalMapper in dynamic mode showed better classification performance than the EM61-Mk2 cart**. The MetalMapper analyses often led to more TOIs correctly classified and/or greater reductions in Non-TOI digs.
- **The MetalMapper in static mode produced better classification performance than in dynamic mode.** Some of the static analyses led to the correct classification of most or all TOIs while reducing Non-TOI digs by over 50%.
- **The TEMTADS outperformed the MetalMapper in static mode.** Most of the TEMTADS analyses led to the correct classification of most or all TOIs while reducing Non-TOI digs by over 90%.
- **Advanced geophysical models led to excellent classification results.** At this challenging site, high-quality data did not consistently lead to excellent results—high-quality data processing was also needed. Use of advanced geophysical models on the MetalMapper and TEMTADS static data led to the correct classification of all TOIs while reducing Non-TOI digs by 92% and 95%, respectively.
- **Second-pass analyses improved classification performance.** A ranked anomaly list was created in the first pass. Ground-truth labels were then provided for those anomalies classified as “dig”. The labels were compiled by digging up the buried targets and identifying them as true TOI or Non-TOI. In the second pass, the ground truth labels were used to refine the classification algorithms and reclassify some anomalies that had originally been classified as “don’t dig.” This mimicked what could occur in a real remediation project, where ground truth could be considered as it becomes available.
- **Commercial geophysics companies performed well.** With appropriate training, non-experts satisfactorily used the UX-Analyze and UXOLab software to process EM61-Mk2 and MetalMapper data. In fact, the commercial geophysics companies outperformed the more experienced organizations that mentored them.

Key Recommendations

- **Future demonstrations should plan time and resources for sufficient quality control.** In this demonstration, quality-control checks promptly caught and addressed problems with data collection and anomaly detection. This is especially important when using dual-mode instruments like the MetalMapper.

- **All classification analyst teams should be given the opportunity to perform multiple-pass analyses.** This will help ESTCP better understand the ground truth feedback processes that could be used in a real remediation project to give stakeholders more confidence in the final don’t dig threshold.
- **Classification analyst teams should be limited to only a few different types of analyses.** In this demonstration, commercial geophysicists with little to no experience in UXO classification outperformed their mentors. This may be because the mentors performed many different types of analyses, spreading their time and resources thin.
- **Commercial geophysics companies should be encouraged to take part in future demonstrations.** This could jump-start technology transfer because the demonstrations provide an excellent opportunity for commercial firms to receive training on UXO classification technologies.

Contents

Acronyms

AUC	Area Under the Curve
DGPS	Differential Global Positioning System
DSB	Defense Science Board
EMI	Electromagnetic Induction
ESTCP	Environmental Security Technology Certification Program
FAR	False-Alarm Rate
FN	False-Negative
FP	False-Positive
GPS	Global Positioning System
HGL	HydroGeoLogic Inc.
IDA	Institute for Defense Analyses
IDL	Interactive Data Language
ISO	Industry Standard Object
IVS	Instrument Verification Strip
NRL	Naval Research Laboratory
Pd	Probability of Detection
Pfa	Probability of False Alarm
RMS	Root Mean Square
ROC	Receiver Operating Characteristic
RTK	Real Time Kinematic
SERDP	Strategic Environmental Research and Development Program
SIG	Signals Innovation Group
SNR	Signal-to-Noise Ratio
TEMTADS	Time-domain Electromagnetic Multi-sensor Towed Array Detection System
TN	True-Negative
TOI	Target of Interest
TP	True-Positive
UXO	Unexploded Ordnance

1. Introduction

Unexploded ordnance (UXO) are explosive, propellant, or chemical-containing munitions that were armed and fired but remain unexploded [27]. UXO continues to pose a risk of detonation, sometimes decades after initial use [27]. Many sites contaminated with UXO are used or intended for civilian purposes, often with no restrictions [24][27]. To eliminate the risk of unintended detonation, the UXO must be identified and removed from these sites, in a process known as "UXO remediation" [24][27].

UXO often becomes buried in the ground and can therefore be impossible to identify by eye. Instruments have been developed to detect buried metallic targets; all detected targets must then be dug up using expensive, safety-oriented procedures in case one or more of them turn out to be UXO. Most targets turn out to be "clutter," however, a term describing nonhazardous items such as fragments from already-exploded munitions, other scrap metal, etc. [24].

In 2003, the Defense Science Board (DSB) released a study on UXO [24]. It found that in the United States alone, over 10 million acres of land are suspected of UXO contamination due to their prior uses as battlegrounds, military test sites, or military training camps. Tens of billions are dollars are necessary to remediate these sites; at current funding rates, this will likely take several decades. In a typical remediation project, as much as 75% of remediation costs are associated with recovering targets that in retrospect could have been left safely in the ground, since they turned out to be clutter. In fact, more than 99% of recovered targets turn out to be false alarms. The DSB pointed out that reducing the false-alarm rate (FAR) from 99% to a lower, yet still relatively high, rate could significantly reduce the costs of UXO remediation. This would allow a larger expanse of land to be cleared with the same amount of funding. Classifying a buried target as either UXO or clutter *before recovery is even attempted* could reduce the false-alarm rate while still ensuring the recovery of most or all hazardous items.

UXO classification technologies have been developed under the Strategic Environmental Research and Development Program (SERDP) and refined under the Environmental Security Technology Certification Program (ESTCP) [23]. Before 2007, testing of these technologies was limited to artificially constructed, standardized test sites such as those at the Aberdeen and Yuma Proving Grounds. Demonstrations carried out at standardized sites have proven to be useful for research and development purposes. However, the results of standardized site demonstrations are not always directly applicable to real remediation projects, as both UXO and clutter are emplaced at the standardized sites according to preconceived notions of which particular types and sizes of targets should be buried at which particular locations, depths, and orientations. Since 2007, then, demonstrations have also been conducted at *live sites* [33][34]. Live sites are locations suspected of UXO contamination due to their previous military uses. The results of live-site demonstrations can be more readily extrapolated to real remediation projects because the results are based on real-world targets under real-world conditions. In particular, live sites can be used to demonstrate the performance of not only individual classification technologies, but also an entire decision-making process that mimics what could occur in a real remediation project. This is crucial for gaining acceptance of classification technology in the UXO community.

ESTCP has designed a series of increasingly challenging live-site demonstrations to validate and gain acceptance of UXO classification technologies. The first demonstration was at the former Camp Sibert, AL, in 2007. Results showed that UXO classification was possible at a site with benign terrain and geology that was contaminated with a single, large munition type. The estimated size of the buried target proved to be the single most discriminating feature between the large UXO and the small clutter [33]. The second demonstration was held in 2009 at the former Camp San Luis Obispo, CA, a site with more challenging terrain and geology, as well as a variety of munition types, all in the medium-to-large size range. Results of this second demonstration showed that with the use of more advanced data-collection instruments, UXO classification was still possible under these more difficult conditions. Once again, size was the most discriminating feature, separating the medium-to-large UXO from the predominantly small clutter [34]. The third demonstration, held in 2010, was designed to be even more challenging. The former Camp Butner, NC, was specifically chosen for its wide variety of munition types, some of which were very small and of a comparable size to clutter. Size was not expected to be a discriminating feature, pushing the participants to explore other features on which classification could instead be based.

One of the two main goals of this third demonstration was to test and validate, on a more challenging site, the instruments and software that had proved successful in the first two demonstrations [2]. To that end, both currently available and emerging electromagnetic induction (EMI) instruments were used to collect data, and both commercially available and custom-built software were used to process and classify the data. The EM61-Mk2 cart was used to collect data in dynamic mode. Built by Geonics Ltd., this instrument consists of one coil that can transmit an electromagnetic field, as well as a second receive coil. The EM61-Mk2 cart has become the standard instrument for UXO remediation. The more advanced MetalMapper, built by Geometrics Inc., was used to collect data in both dynamic and static modes. This instrument consists of three orthogonal transmit coils as well as seven triaxial receive coils. Similarly, the Time-domain Electromagnetic Multi-Sensor Towed Array Detection System (TEMTADS) consists of a 5 × 5 array of transmit and receive coils. This instrument was developed by the Naval Research Laboratory (NRL) and was used to collect static data. Several different software packages were used to process the collected data, including the commercially available UX-Analyze module to Oasis montaj, the UXOLab software developed by the University of British Columbia, and other custom-built software packages.

The second goal of this demonstration was to involve the regulatory community in the design, implementation, and evaluation of the demonstration in an effort to better understand what might be required in a real remediation project if buried targets were classified as clutter and therefore left in the ground [2]. To that end, state regulators, representatives of the Environmental Protection Agency, members of the U.S. Army Corps of Engineers, and other stakeholders were invited to participate in an Advisory Group. This group strongly influenced the design, execution, and evaluation of the demonstration, as had been done for the previous two demonstrations.

The Institute for Defense Analyses (IDA) was assigned responsibility for assisting with the design, execution, and evaluation of the demonstration under a task entitled "ESTCP/SERDP: Assessment of Traditional and Emerging Approaches to the Detection and Identification of Surface and Buried Unexploded Ordnance." In support of ESTCP, IDA created a protocol for seeding inert UXO in the demonstration area at the former Camp Butner, as well as a second protocol for scoring the final classification deliverables against ground truth [31][32]. Finally, IDA scored the 54 separate deliverables submitted by the 8 classification analyst teams: CH2M HILL, Dartmouth College, Geometrics Inc., NAEVA Geophysics Inc., Parsons Inc., SAIC, Signals Innovation Group (SIG), and Sky Research Inc. This comprehensive report describes the demonstration in detail and serves as a complement to the more concise report produced by ESTCP [29].

The following sections describe the site preparations and data-collection procedures used at the former Camp Butner. The methods used to classify the collected data and the process used to score the classification deliverables are also described. The final sections present the results of the scoring, summarize the findings of the demonstration, and make recommendations for future demonstrations.

2. Site Preparation

Careful preparations were made for the third live-site UXO classification demonstration. This section describes the steps that were taken before data collection, including the methods used to select a site for the study, select a particular section of the site for the demonstration area, and seed the demonstration area with inert UXO.

A. Selecting a Site

The former Camp Butner was selected for its challenging variety of munitions. This site had been used as a military training camp during World War II [19]. Historical records showed that multiple munition types had been fired at the site, ranging from large 105 and 155 mm projectiles to small 37 mm projectiles comparable in size to clutter [19]. Based on input from the Advisory Group, ESTCP intentionally chose a site with such small UXO in order to challenge the classification technologies, pushing the demonstration participants to base their classifications on features other than the estimated size of the buried target. Figure 1 shows an aerial photograph of the site, and Figure 2 shows a photograph from the ground. The former Camp Butner is grassy and relatively flat.

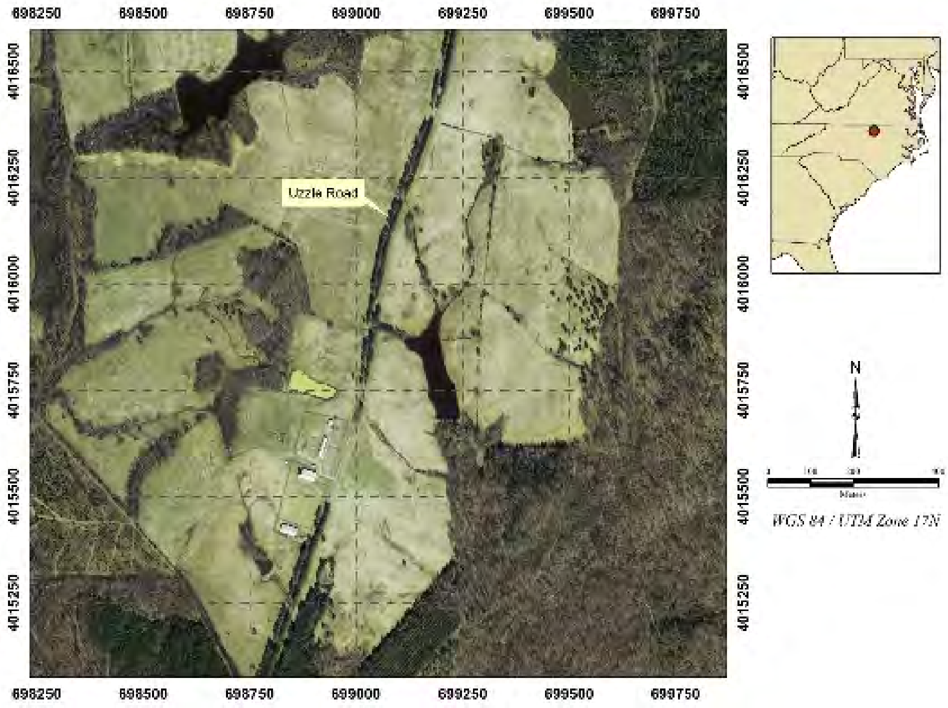


Figure 1: An aerial photograph of the former Camp Butner. Taken from [29].

Figure 2: A photograph of the former Camp Butner. Taken from [29].

B. Selecting the Demonstration Area

Transects were taken to select sections of the former Camp Butner that were potentially suitable for the demonstration. As shown in Figure 3, HydroGeoLogic Inc. (HGL) used an EM61-Mk2 sensor to collect transects over four separate sections of the site, with a 20 m separation between transect lines [8]. Peaks were identified in the collected data. A threshold of 20 mV was applied to the sum of the four sensor channels, each channel representing the signal received by the sensor at one of four points in time after exciting the buried target with an electromagnetic field pulse. Each peak above threshold was considered an "anomaly" with respect to background. The anomaly densities were calculated for each section of the site, and 24 specific acres were chosen for further investigation.

An initial survey was then done to select the final demonstration area. Under subcontract to HGL, NAEVA Geophysics Inc. collected survey data over the 24 acres selected from the transects [7]. An EM61-Mk2 sensor was used here as well, this time with a line spacing of 0.5 m. Figure 4 shows another aerial photograph of the site, this time overlaid with a map of the EM61-Mk2 survey data. Anomalies were identified by applying a threshold of 5 mV to the second sensor channel, as this channel often shows the largest signal-to-noise ratio (SNR). The site was separated into 30 m × 30 m grids, and the anomaly density was calculated for each grid. ESTCP then visually inspected the data map along with the anomaly densities. A total of 20 contiguous grids (4.4 acres in total) in the northeast section of the site were selected for the demonstration area. This area is outlined in blue in Figure 4. These 20 grids exhibited anomaly densities high enough to sufficiently challenge the classification methods but low enough to allow the recovery of all buried targets to use as ground truth in scoring.

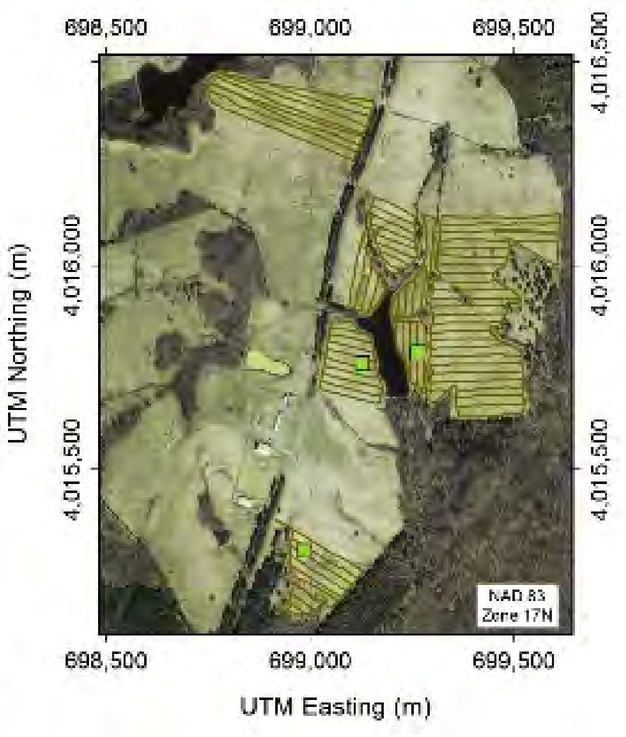


Figure 3: An aerial photograph of the former Camp Butner. Transect lines are marked in brown. Green squares mark potential grids for intrusive investigation. Taken from [29].

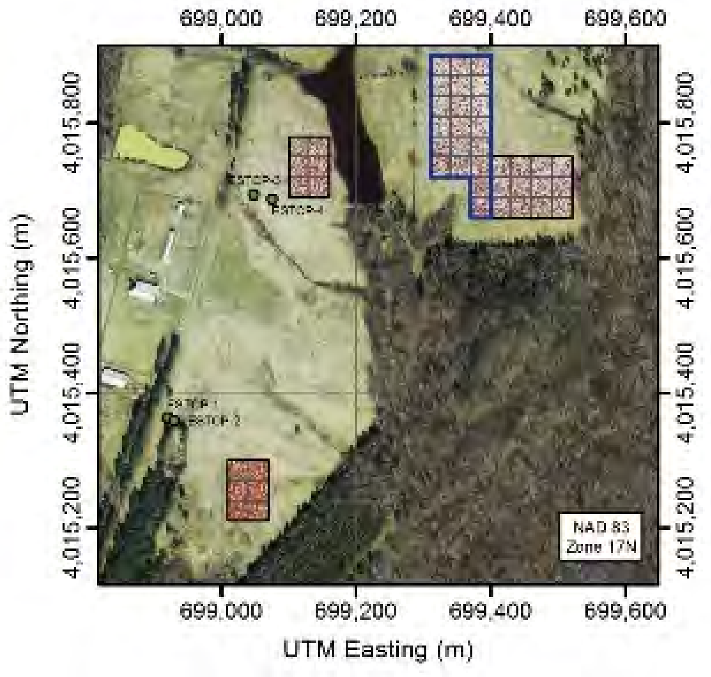


Figure 4: An aerial photograph of the former Camp Butner. The initial EM61-Mk2 data map (lower coil, second time gate) is overlaid before seeding. The selected demonstration area is outlined in blue in the northeast section of the site. Taken from [29].

C. Seeding the Demonstration Area

The purpose of conducting a UXO classification demonstration on a *live site* is to make use of the site's native targets. In contrast, standardized sites, such as the Aberdeen and Yuma Proving Grounds, are first cleared of all native targets and then seeded with preselected UXO and clutter before a demonstration begins. Although standardized sites allow more control over the demonstration, the results can be difficult to extrapolate to a real remediation project. In contrast, live-site demonstrations have the credibility of mimicking the real world—classification results are based on real-world targets native to real-world sites.

Native clutter items are common at live sites, but native UXO items are rare. For example, thousands of clutter items were found at both the former Camp Sibert and the former Camp San Luis Obispo [33][34]. This led to a high statistical confidence in the demonstration results, which were that approximately 30%–50% of anomalies produced by native clutter items could be correctly classified. This was not the case for native UXO, however. At the former Camp San Luis Obispo, only 44 native UXO were found in the 10 acres assigned to the demonstration area [34]. Furthermore, zero native UXO were found in the former Camp Sibert's demonstration area [33].

A demonstration cannot assess, with confidence, the ability to correctly classify anomalies produced by UXO if there is not an adequate number of UXO to begin with. Therefore, inert UXO were seeded at all three demonstration sites to provide statistical confidence in the classification performance metrics. The locations and identities of the seeded UXO were kept hidden from the demonstration participants until after scoring was complete. In contrast, no clutter items were seeded, since a very large number of clutter items were already native to the site.

The reliance on seeded UXO was the main limitation of this series of demonstrations. Ideally, the demonstration area would have been sufficiently large such that valid statistics could have been calculated from only UXO that were native to the site. Such a large area would have also included an extremely large number of native clutter items, however. In this type of demonstration, all targets must be recovered to constitute ground truth for scoring. Yet the cost of recovering all targets from such a large area would have been prohibitively high. Budget constraints limited the size of the demonstration areas at the former Camp Sibert, former Camp San Luis Obispo, and former Camp Butner. This resulted in few recovered targets that turned out to be native UXO and so warranted the need for seeded UXO. Future demonstrations will also have limited budgets, and so future demonstration areas will also be limited in size, leading to the recovery of few native UXO. Therefore, the seeding of UXO will remain necessary.

The seeding of UXO influenced only some of the metrics calculated in this series of demonstrations. Only those metrics that were based on UXO were influenced, such as the number or percentage of UXO that could be correctly classified and must therefore be recovered in a real remediation project. In contrast, the seeded UXO did *not* influence those metrics that were based on the clutter items, such as the number or percentage of clutter items that were *incorrectly* classified and must therefore be *unnecessarily* recovered in a real remediation project—that is, the metrics describing the false alarms.

1. Selecting UXO for Seeding

ESTCP gave considerable thought for choosing the types of UXO to seed at the former Camp Butner. To keep the demonstration as close as possible to a real-world situation, the desired seeds must be as similar as possible to the UXO previously fired at the site. Historical records showed that several types of munitions had been fired at the former Camp Butner, including 37, 105, and 155 mm projectiles [19].

An intrusive investigation was performed over a small section of the site to confirm the historical records [21]. HGL, together with Ordnance Explosive Remediation Inc. (a commercial UXO remediation company), recovered all metallic objects within a 100 ft × 65 ft area of the southernmost section of the site, the southernmost green square in Figure 3. A total of 593 items were recovered, all consisting of debris from previously exploded UXO or other metallic items. No items were found deeper than 50 cm (20 in). The most common UXO debris was from 105 and 155 mm projectiles, along with some fragments of 37 mm projectiles.

Based on the historical records and the results of the intrusive investigation, ESTCP decided to seed three types of UXO at the former Camp Butner: 37 mm projectiles, 105 mm projectiles, and M48 fuzes. 155 mm projectiles were not seeded because their extremely large size made them too easy a target for UXO classification. On the other hand, 37 mm projectiles and M48 fuzes (fuzes detached from their parent 105 mm projectiles) were small enough to challenge UXO classification methods. The Advisory Group noted that 105 mm projectiles are often found with their M48 fuzes detached and lying nearby; the detached fuzes remain filled with explosive material and can still pose a risk of detonation.

ESTCP queried munitions stores across the United States for inert munitions of these three types. A total of 112 inert 37 mm projectiles were accumulated from different sources, along with 27 inert 105 mm projectiles. Because few detached M48 fuzes could be found, ESTCP contracted with the Naval Research Laboratory (NRL) to manufacture 24 metallic objects of the same size, shape, material composition, and wall thickness as an M48 fuze.

With the input of the Advisory Group, ESTCP also determined the depths of interest for the seeded UXO. These were based on the maximum depth at which each UXO type might be typically found, along with the site conditions specific to the former Camp

Butner that might affect the likely depths. Table 1 provides the depths of interest used for the UXO types expected at Butner.

Table 1: Depths of interest for the UXO types expected at the former Camp Butner.

UXO	Depth of Interest
105 mm projectile	60 cm (2 ft)
37 mm projectile	30 cm (1 ft)
M48 fuze	30 cm (1 ft)

2. Creating the Seed Plan

The seed plan consisted of a list of locations for seeding the UXO, along with instructions for interpreting the list [32]. The list was separated into three sections: (1) the demonstration area, (2) the instrument verification strip (IVS), and (3) the training pit. The first 160 UXO were seeded in the demonstration area, including 110 inert 37 mm projectiles, 26 inert 105 mm projectiles, and 24 inert M48 fuzes. The purpose of these seeds was to guarantee the existence of a large number of UXO to ensure sufficient statistical confidence in the classification performance metrics. The second, much shorter section listed six targets for the IVS: two 37 mm projectiles, two spherical shot puts, and two 1 in × 4 in pipe nipples called "Industry Standard Objects (ISOs)" that were approximately the same size and shape as 37 mm projectiles. These six targets were emplaced in the IVS after clearing the IVS of all other targets. The purpose of the IVS was to allow the data-collection teams to recalibrate their instruments on a daily basis using known targets. The final section listed the location of the training pit. Data-collection teams used the training pit to collect data from expected UXO types to learn their expected signatures at different depths and orientations. To that end, one 37 mm projectile, one 105 mm projectile, and one M48 fuze were reserved for data collection in the training pit. (Section 2.c.3 details deviations made to the original seed plan.)

a. Demonstration Area

IDA selected the intended locations of all 160 UXO to be seeded in the demonstration area [32]. First, the thresholded EM61-Mk2 data map was visually examined. It was assumed that all anomalies in the map were produced by native targets or geology. The 160 selected seed locations were far from each other and far from any native anomaly. When selecting the locations of the seeded UXO, anomalies were avoided since multiple, closely spaced targets, such as a UXO seeded next to a piece of native scrap metal, had generally been difficult to separate and classify in previous classification studies [34]. Figure 5 is a close-up view of one 30 m × 30 m grid of the thresholded EM61-Mk2 data map. Colored circles indicate the intended seed locations; all were far from each other and any native anomaly.

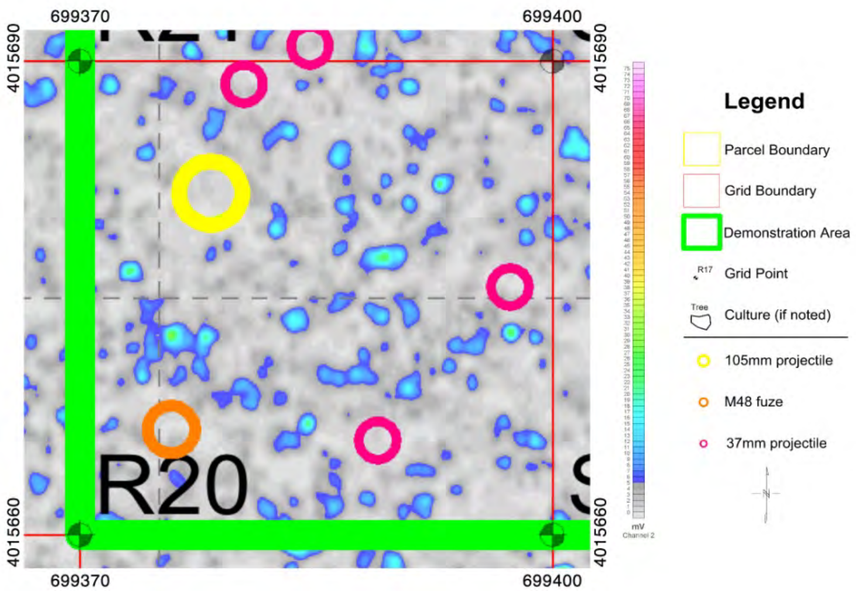


Figure 5: A close-up view of one 30 m × 30 m grid in the initial EM61-Mk2 data map (lower coil, second time gate) before seeding. Native anomalies exceeding the threshold of 5 mV are shaded in color, and the background is shaded in gray. Pink, orange, and yellow circles of radius 1.2 m, 1.5 m, and 2.0 m mark the intended locations of 37 mm projectiles, 105 mm projectiles, and M48 fuzes, respectively. All intended seed locations are far from each other and any native anomaly. Taken from [32].

Analysts at SAIC assisted in determining the minimum spacing between intended seed locations [40][41]. Using a geophysical model, the analysts estimated the signal that would be measured by the EM61-Mk2 from each expected UXO buried at its depth of interest and each of three orthogonal orientations. Then, they traced a 3.0 mV contour in the estimated data and calculated the diameter of the contour. The left side of Figure 6 shows the contours traced for a 105 mm projectile buried at its 60 cm (2 ft) depth of interest and in three different orientations. The average diameter of the contours was 3.07 m. Similar plots are shown for a 37 mm projectile and M48 fuze, both buried at their 30 cm depths of interest, with average contour diameters of 1.80 m and 2.20 m, respectively. IDA applied a 33% safety margin to the estimated contour diameters and then used these final numbers to determine the minimum spacing between seeded UXO. Specifically, all locations assigned to 105 mm projectiles were spaced at least 4.0 m away from native anomalies and other seed locations. All locations assigned to 37 mm projectiles were spaced at least 2.4 m away, and all locations assigned to M48 fuzes were spaced at least 3.0 m away.

IDA randomly assigned each seeded UXO to a specific depth (up to the depths of interest), inclination angle, and azimuth angle [32]. For example, as 30 cm was the depth of interest for 37 mm projectiles, all seeded 37 mm projectiles were randomly assigned to depths ranging from 10 to 30 cm. M48 fuzes were also randomly assigned to depths of 10 to 30 cm, because 30 cm was their depth of interest as well. The 105 mm projectiles were randomly assigned to depths of 20 to 60 cm, as their depth of interest was 60 cm. Inclination angles were assigned by considering the Advisory Group's comments that in real remediation projects, most UXO of these types are found in horizontal orientations. Therefore, most of the seeded UXO at the former Camp Butner were assigned inclination angles within 45 degrees of horizontal. Finally, the seeded UXO were randomly assigned to different azimuth angles with respect to north.

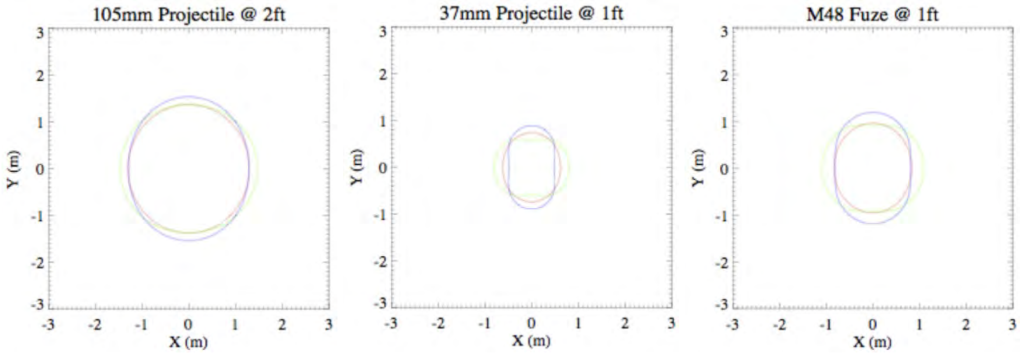


Figure 6: 3 mV contours traced along the estimated EM61-Mk2 signal amplitudes (lower coil, second time gate) for different UXO types seeded at their depths of interest, including: (left) a 105 mm projectile at 60 cm (2 ft), (middle) a 37 mm projectile at 30 cm (1 ft), and (right) an M48 fuze at 30 cm (1 ft). Contours were estimated for each of three orthogonal orientations. Taken from [40].

The seed plan instructed the emplacement team to bury each UXO as close as possible to its intended location, depth, inclination, and azimuth [32]. The plan did allow for minor deviations if needed. For example, the emplacement team was instructed to survey the area around an intended location with a hand-held EMI detection device. The purpose of this step was to check for native targets or geology that for any reason had failed to produce an anomaly on the initial EM61-Mk2 data map. If no anomalies were detected, then the emplacement team was instructed to proceed with seeding the UXO. On the other hand, if an anomaly was detected, then the emplacement team was instructed to choose a nearby location to seed the UXO. In another example of a deviation from plan, the emplacement team was instructed to alter the depth and orientation of a seeded UXO if its intended burial parameters did not allow at least 10 cm of dirt over the top of the buried seed.

b. Instrument Verification Strip

The emplacement team was instructed to manually select the intended locations for the six targets seeded in the IVS [32]. The main purpose of the IVS was to recalibrate the instruments on a daily basis. ESTCP recommended situating the IVS directly west of the demonstration area. This area exhibited few anomalies representing native targets or geology. Furthermore, the data-collection teams could easily access this area with their instruments at the beginning and ending of each day for calibration. Based on this guidance, a strip of land was chosen that was parallel to and 5 m away from the western boundary of the demonstration area [32]. The emplacement team was instructed to first search for and clear any native metallic items from this area before seeding the targets 5 m apart from each other.

Target depths, inclinations, and azimuths were selected for each target seeded in the IVS [32]. Both the 37 mm projectiles and the ISOs (1 in × 4 in pipe nipples roughly the same size and shape as 37 mm projectiles) were assigned horizontal cross-track orientations, a directly horizontal inclination angle with an azimuth angle perpendicular to the length of the strip. This orientation was chosen to allow instrument calibration in a worst-case scenario—horizontal cross-track generally provides the weakest signal for EMI sensors. One 37 mm projectile and one ISO were assigned a depth of 15 cm, with the other of each type assigned to 30 cm. The two shot puts were assigned depths of 45 cm; orientation was irrelevant because these were spheres.

c. Training Pit

The seed plan also instructed the emplacement team to dig a training pit 8 m north of the IVS [32]. First, this area was to be cleared of all metallic items within a 5 m radius. Then, a pit was to be dug in the center of the cleared area, 0.50 m in radius and 0.60 m in depth. One 37 mm projectile, one 105 mm projectile, and one M48 fuze were reserved so that data could be collected from these UXO at different depths and orientations requested by the classification analyst teams. These data could later assist the teams in learning the expected signatures of the UXO, such that better classification decisions could then be made within the demonstration area.

3. Seeding the UXO

UXO were seeded at the site according to the seed plan. First, HGL performed a surface clearance to remove any metallic items resting on the ground [20]. Then, Parsons Inc., a commercial geophysics company, buried the UXO at or near the intended locations, depths, and orientation angles.

Some deviations from the plan were made. In the demonstration area, the plan specified one more M48 fuze than was available. Therefore, Parsons removed the fuze from one of the 105 mm projectiles and used that instead. In the IVS, the plan specified one more 37 mm projectile than was available. Therefore, Parsons used a third ISO in place of the deepest intended 37 mm projectile.

After placing each target in the ground, but before covering it with dirt, Parsons recorded several pieces of information:

- The identification number of the seeded target.

- The type of the target (e.g., “37 mm projectile,” “105 mm projectile,” “M48 fuze”).
- The easting, northing, and depth coordinates for the nose, center, and tail of the target, with depth measured with respect to the average of one or more surveyed points on the lip of the hole.
- The azimuth and inclination angles of the target.
- A photograph of the target, with the identification number clearly written on the target and a ruler clearly placed next to the target.

The emplacement team also completed final preparations to the site. They replaced dirt in the holes and leveled the final burial locations of the seeds. They also attempted to replace the grass plug over the burial locations. Several weeks then passed before data collection began.

3. Collecting Data

Two types of data can be collected for UXO remediation: dynamic data and static data. Dynamic data are collected at a relatively steady rate as the data-collection instrument travels over the ground. The main purpose of collecting dynamic data is to detect individual anomalies indicating buried targets. In some cases, dynamic data can also be used to classify the buried targets into two groups: those that must be recovered (likely UXO) and those that may remain in the ground (likely clutter). In contrast, static data are collected at one particular location at a time as the data-collection instrument remains at rest. The locations must be known in advance; they are typically the locations of the anomalies detected in previously collected dynamic data. In general, static data have a higher resolution and SNR than dynamic data and can therefore be much more useful for UXO classification.

Both types of data were collected at the former Camp Butner, once seeding was complete. This section describes the instruments and methods used to collect dynamic data, the methods used to detect anomalies in the dynamic data, the results of scoring the detected anomalies against the seeded UXO, and the instruments and methods used to collect static data at the locations of the detected anomalies. Table 2 summarizes the three EMI instruments used for data collection at the former Camp Butner.

Table 2: Data collection instruments at the former Camp Butner.

Instrument	Sensor	Mode	Status
EM61-Mk2 cart	Standard transmit/receive coil	Dynamic	Industry standard
MetalMapper	Three orthogonal transmit coils with seven triaxial receive coils	Dynamic and static (dual-mode, self-cued)	Emerging
TEMTADS	5 × 5 array of transmit/receive coils	Static	Emerging

A. Collecting Dynamic Data

Two EMI instruments collected dynamic data at the former Camp Butner: the EM61-Mk2 (the same instrument used to collect transects and the initial survey data before seeding) and the MetalMapper. Each instrument is described briefly below. More detail can be found in the documents written by the data-collection teams [3][47][48][53][54][60][61].

1. EM61-Mk2 Cart

The EM61-Mk2 cart is the most commonly used EMI instrument for UXO remediation. It consists of a standard EM61-Mk2 sensor mounted on a two-wheeled cart. Sold by Geonics Ltd. since the 1990s, the sensor consists of a 1 m × 0.5 m receive coil mounted 30 cm above another similar sized structure containing both a transmit coil and second receive coil. As shown in Figure 7, current passing through the lower coil transmits an electromagnetic field. Changes in this primary field, such as when the current is turned on and off, induce eddy currents in the buried target, a process known as "illuminating" the target. The eddy currents give rise to a secondary electromagnetic field that, in turn, induces a secondary current through both the upper and lower coils of the sensor. The secondary current induced in each coil passes through a known resistance, resulting in a measurable change in voltage. The EM61-Mk2 cart can be configured to measure either (1) the voltage change in the lower coil at four time gates (geometrically spaced in time from 216 μs to 1.3 ms) immediately after the primary current is turned off or (2) the voltage change in the upper coil at the first time gate and in the lower coil at three time gates. The operator of the instrument wears a backpack with a battery and other electronics [17][48].

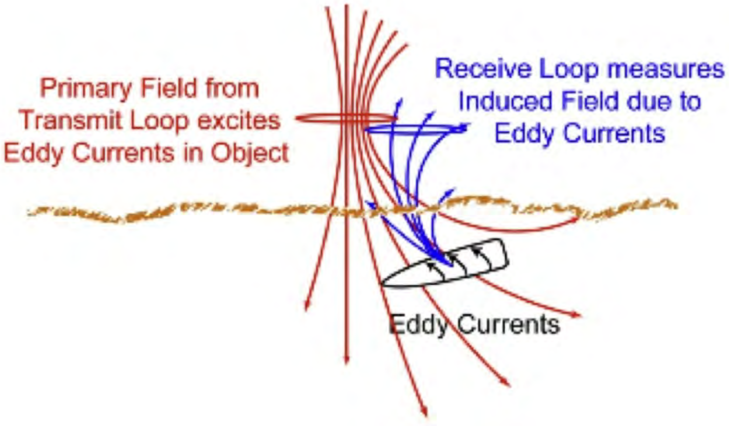


Figure 7: The EM61-Mk2 sensor consists of a lower coil that transmits a primary electromagnetic field. Changes in the primary field induce eddy currents in the buried target. The eddy currents give rise to a secondary electromagnetic field. At the former Camp Butner, the strength of the secondary field through a second lower coil was measured at four time gates after the primary field was turned off. Taken from [48].

NAEVA Geophysics Inc. operated the EM61-Mk2 cart at the former Camp Butner [47][48], as shown in Figure 8. NAEVA configured the instrument to measure the strength of the secondary field through the lower receive coil at four time gates. This was done to provide the maximum temporal extent for assessing the decay of the secondary field over time because this decay is known to be a discriminating feature in UXO classification [33][34]. NAEVA used the IVS to calibrate the sensor at the beginning and end of each day. In the demonstration area, the NAEVA operator pulled the cart over the ground in straight lines that were 0.5 m apart. Although UXO remediation projects often use a line spacing as wide as 1.0 m, NAEVA used the closer line spacing to enable the collection of high-resolution data that could support classification. Finally, NAEVA used the training pit to collect data from the expected UXO at different depths and orientations, as requested by the classification analyst teams. In all cases, the NAEVA operator used a Trimble 5700 real time kinematic differential global positioning system (RTK DGPS) to track the position of the sensor as the cart was pulled over the ground.

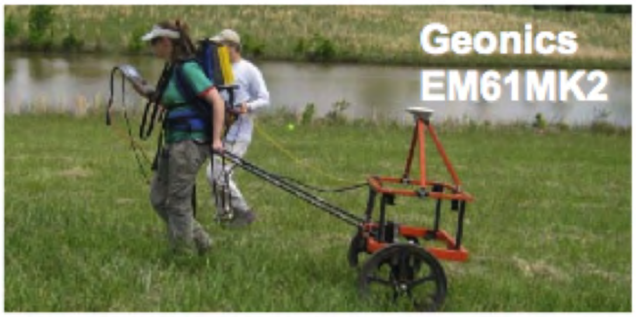


Figure 8: NAEVA Geophysics Inc. operated the EM61-Mk2 cart at the former Camp Butner. Taken from [51].

The EM61-Mk2 cart has both advantages and disadvantages for UXO classification. On the positive side, this instrument is inexpensive compared with more advanced systems. Furthermore, the instrument is well known to commercial companies involved in UXO remediation; many commercial geophysicists know how to operate the instrument and analyze its data. On the negative side, it uses a mono-static sensor, meaning that it measures the strength of the secondary field in only one direction (the direction through the receive coil). However, to estimate the characteristics of a target buried at a particular spot in the ground, the analyst must consider the strength of the secondary field in *all* directions. To do this, the analyst must patch together data collected as the cart was pulled directly over the target (where the receive coil is directly above the target), as well as over separate, nearby lines (where the receive coil is above and to the side of the target). Inaccurate measurements of the cart's position from line to line can introduce errors into this process, resulting in a mischaracterization of the buried target. Another limitation of the EM61-Mk2 cart is that it measures the strength of the secondary field at only four time gates, the latest occurring only 1.3 ms after the primary field is turned off, a much shorter span of time in comparison to the advanced instruments. Thus the secondary field's decay over time is more difficult to assess, leading to further difficulties in characterizing the buried target.

2. MetalMapper

The MetalMapper is an advanced EMI instrument developed by Geometrics Inc. under ESTCP funding. As shown in Figure 9, it consists of three 1 m × 1 m orthogonal transmit coils, as well as seven triaxial receive coils. A case contains the battery and other electronics, including a RTK DGPS and an inertial measurement unit to resolve the sensor's yaw, pitch, and roll. The MetalMapper can be operated in both dynamic and static modes [53][54][60].

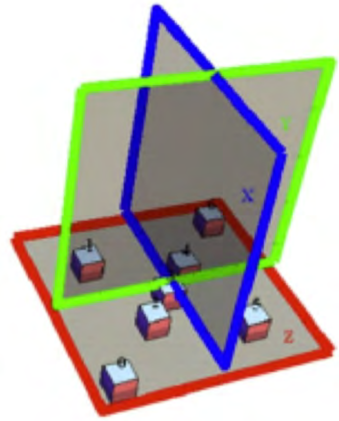


Figure 9: The MetalMapper consists of three orthogonal transmit coils, as well as seven triaxial receive coils. Only one transmit coil is used in dynamic mode, with all seven receive coils used to sense the secondary electromagnetic field. In static mode, all transmit coils and all receive coils are used. Taken from [29].

In dynamic mode, the MetalMapper functions slightly differently from the EM61-Mk2 cart. Current is passed through only the horizontal transmit coil, creating a vertically oriented primary field that, as with the EM61-Mk2 cart, illuminates the buried target in only one direction. Unlike the cart, however, the MetalMapper senses the secondary field on each of the seven triaxial receive coils, resulting in 21 separate measurements. The voltage change in each axis of each receive coil is measured at different time gates after the primary field is turned off; the operator can specify the exact time gates used [53].

Geometrics collected MetalMapper dynamic data at the former Camp Butner. Its subcontractor, Sky Research Inc., also collected dynamic data with a second, newer system in the northwest portion of the demonstration area. Both organizations configured their time gates from 24 μs to 904 μs and used a line spacing of 0.75 m. Figure 10 shows the MetalMapper deployed on the front lift of a Kubota tractor. Like NAEVA, both organizations used the IVS to calibrate their instruments at the beginning and end of each day. In addition, the team collected the training pit data requested by the classification analyst teams [53].

There are advantages and disadvantages to collecting dynamic data with the MetalMapper. The main advantage is that the MetalMapper is a multi-static sensor: at any particular time gate, the strength of the secondary field is sensed on seven triaxial receive coils. Because each receive coil actually consists of three separate orthogonal coils, each coil can measure the strength of the secondary field in three orthogonal directions. Furthermore, the seven receive coils are mounted at a fixed distance from each other, leading to a very high relative position accuracy of the seven locations at which the field is measured. All of this reduces errors in the characterization of the buried target. On the other hand, the latest time gate was configured to only 904 μs in this demonstration, even earlier than the latest time gate of the EM61-Mk2 cart (1.3 ms). Thus, it was even more difficult to assess the decay of the secondary field using the MetalMapper in dynamic mode than with the EM61-Mk2 cart; these difficulties may have degraded the characterization of the buried target. Finally, the MetalMapper is a new instrument and fewer people are familiar with its operation.

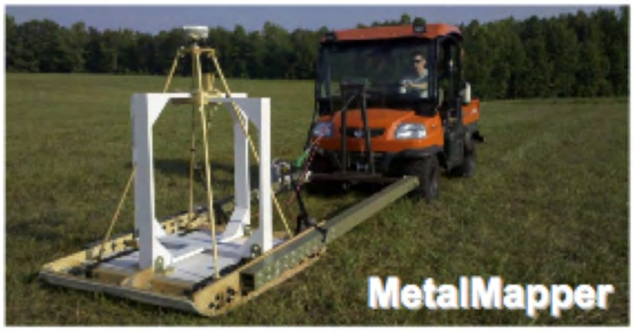


Figure 10: Geometrics Inc. and Sky Research Inc. operated the MetalMapper at the former Camp Butner. Taken from [51].

B. Detecting Anomalies

The primary purpose of collecting dynamic data in a UXO remediation project is to detect anomalies indicating buried targets. To that end, each data collection team created a map of the dynamic data collected at the former Camp Butner. NAEVA plotted the signal (the change in voltage) measured in the lower coil of the EM61-Mk2 cart at the second time gate, as was done before seeding [47][48]. Figure 11 shows NAEVA's dynamic data map for the EM61-Mk2 cart after seeding, superimposed on an aerial photograph of the site. Similarly, Geometrics created a similar map for the MetalMapper dynamic data by plotting a combination of the signals measured in the MetalMapper's seven different receive coils at different time gates [53]. Each team then applied a threshold to its map to detect individual anomalies.

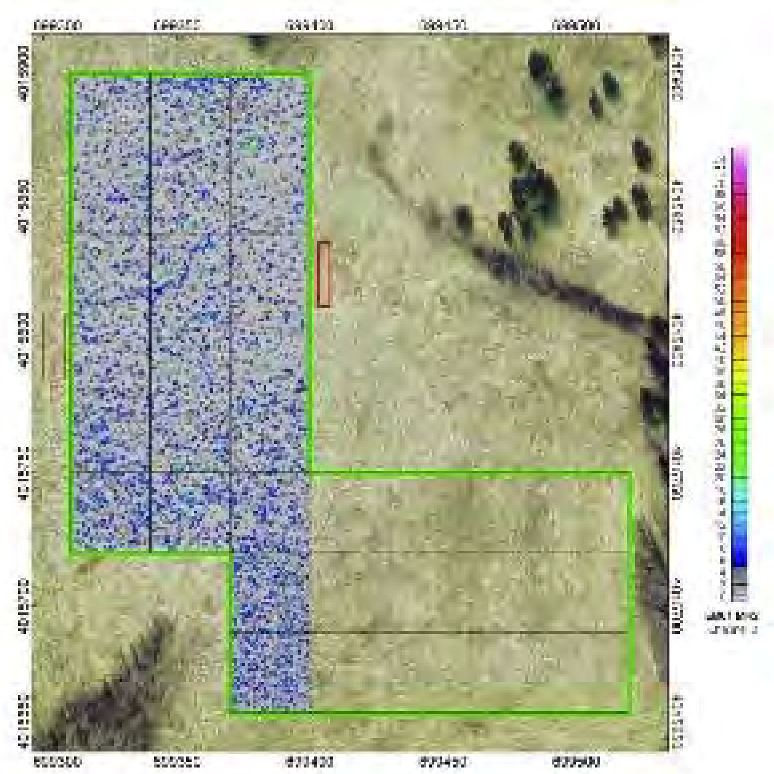


Figure 11: An aerial photograph of the former Camp Butner. The final EM61-Mk2 data map (lower coil, second time gate) is overlaid after seeding. Areas exceeding the detection threshold of 5.2 mV are shaded in color, while all areas below threshold are shaded in gray. Taken from [47].

Both data-collection teams used similar methods to choose their detection thresholds. Using a geophysical model, NAEVA estimated the amplitude of the signal that would be sensed by the EM61-Mk2 cart (lower coil, second time gate) from different UXO buried at different depths and orientations [47][48]. The detection threshold was then set at 5.2 mV, the estimated signal amplitude sensed from a 37 mm projectile (the smallest expected UXO) at its 30 cm depth of interest and its least favorable orientation (that which would lead to the lowest signal amplitude). For example, the red curve in Figure 12 plots the estimated signal amplitude versus depth for a standard 37 mm projectile in the horizontal cross-track orientation. An orange × marks the detection threshold, the signal expected from the 37 mm projectile at its depth of interest of 30 cm and its least favorable orientation. Based on the geophysical model, it was estimated that other expected UXO buried at their maximum depths of interest and least favorable orientations would produce signal amplitudes exceeding the detection threshold. Furthermore, the root-mean-square (RMS) noise floor of the site was much lower than the detection threshold, making it unlikely that any anomalies detected with this threshold would have been caused by noise alone. Geometrics used a similar technique for the MetalMapper, although it applied a 50% safety margin to reduce the original detection threshold from 4.0 mV down to 2.0 mV [53]. The final MetalMapper detection threshold was much closer to the noise floor than that of the EM61-Mk2, making noisy detections more likely in the MetalMapper dynamic data than in the EM61-Mk2.

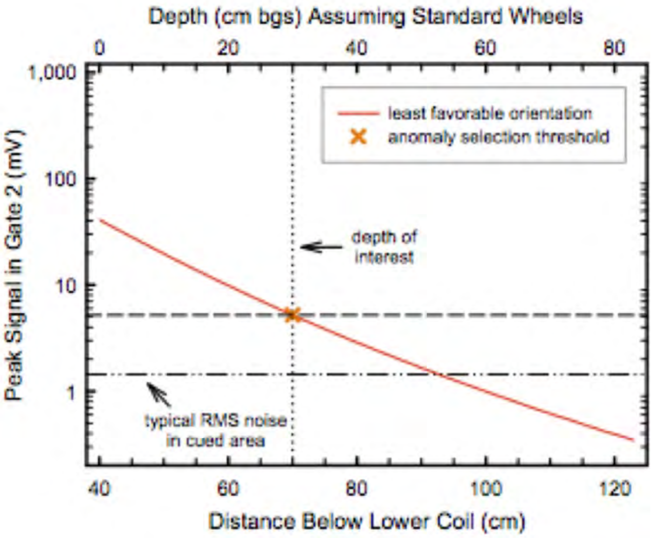


Figure 12: Detection curve for the EM61-Mk2 cart. An orange × marks the EM61-Mk2 detection threshold (5.2 mV) at the former Camp Butner, the estimated amplitude of the signal sensed from a 37 mm projectile (the smallest expected UXO) at its 30 cm depth of interest (70 cm below the sensor platform) and its least favorable orientation. The noise floor in the demonstration area is well below this detection threshold. Taken from [29].

NAEVA used a two-step process for applying its detection threshold to the EM61-Mk2 data map. First, the data was gridded in Oasis montaj, a commercially available software package sold by Geosoft. The software's *gridpeak* function was used to identify peaks in the gridded data exceeding the detection threshold. No smoothing filters were applied to the gridded data before the peaks were identified. Next, NAEVA focused on each pair of peaks that were closer than 0.60 cm (2 ft) from each other. For each pair of peaks, the analyst determined if it was likely that (1) a single buried target produced both peaks (in which case only the highest peak was retained on the EM61-Mk2 anomaly list) or (2) more than one buried target produced the peaks (in which case both peaks were retained on the list). A total of 2304 peaks were listed on the final EM61-Mk2 anomaly list. Each peak was intended to represent one distinct anomaly likely produced by one buried target.

Geometrics used a similar two-step process for detecting anomalies in the MetalMapper dynamic data. In the first step, Oasis montaj was used to automatically identify over 5000 peaks above the final detection threshold (2.0 mV, taking into account the 50% safety margin). In the second step, the analyst inspected the data surrounding each pair of peaks closer than 0.60 m to each other and determined which pairs were likely produced by one buried target. Geometrics reported that this step was difficult due to the large number of closely spaced pairs [53]. In the end, a total of 3765 peaks were listed on the final MetalMapper anomaly list. Many more anomalies were detected in the MetalMapper data (3765) than in the EM61-Mk2 data (2304), likely due to the 50% safety margin that Geometrics applied to its detection threshold, lowering it closer to the noise floor. As stated in the Geometrics data-collection report [53], had a safety margin *not* been employed, roughly half as many anomalies would have been detected.

C. Scoring the Detected Anomalies

The primary purpose of this demonstration was to assess the classification performance of different instrument/algorithm combinations. However, it was also important to assess the *detection* performance of the different instruments alone. Detection results in an anomaly list. In this series of demonstrations, classification resulted in a *ranked* anomaly list, with the detected anomalies ordered according to their estimated likelihood of being clutter. In short, the output of detection is the input to classification; one cannot classify a buried target until or unless it has been detected. (In fact, one cannot even collect static data from a buried target until it has been detected.) To achieve buy-in from the UXO community, then, one must ensure a reasonable detection performance before classification can even be considered. Therefore, early in the demonstration, IDA performed a quick assessment of the detection capabilities of the two dynamic instruments.

Two metrics were estimated for each instrument, the probability of detection (Pd) and the false-alarm rate (FAR). Pd gives the percentage of UXO that were detected and FAR gives the number of anomalies per unit area that did *not* detect any UXO. Each threshold crossing that might represent a UXO should be analyzed during classification, and that is where the ultimate number of false alarms is set. However, as the detection threshold is lowered toward the noise floor, the number of detected anomalies will increase rapidly. Hence, for our purposes, the relative FAR of two sensors at the detection stage is a measure of the margin above noise each sensor has against the smallest signal of interest. For single-pass detection and classification based upon dynamic data only, this margin is critical to classification success, as accurate classification requires high SNR. However, even where classification is based upon static data, higher SNR at the detection threshold produces shorter anomaly lists, thereby reducing the costs to acquire and analyze the static data.

Ideally, full ground truth must be known so that one can determine what percentage of seeded *and native* UXO were detected (Pd) and how many anomalies per unit area did not detect either a seeded *or native* UXO (FAR). Full ground truth was not yet known at this point in the demonstration, though, since none of the detected anomalies had been excavated and so none of the buried targets had been recovered from the ground.

Although the locations of the seeded UXO were known, the locations of any native UXO were not known. Therefore only *estimates* of Pd and FAR could be calculated at this point in the demonstration, based on the seeded UXO only. It was likely that these estimates would turn out to be fairly accurate, though, since native UXO are so rare.

A two-step method was used to estimate Pd for the EM61-Mk2 cart. First, the distance between each seed and its closest EM61-Mk2 anomaly was measured. Table 3 summarizes these measured distances, and Figure 13 is a histogram of the distances. Second, these distances were assessed to determine whether any exceeded 0.60 m. The Advisory Group explained that in real remediation projects, UXO technicians reacquire the location of an anomaly with a global positioning system (GPS) unit and then often use a hand-held magnetometer or EMI instrument to investigate an area within approximately 0.60 m (2 ft) around that location. This allows the technicians to better pinpoint the spot on the ground at which they should dig to recover the buried target. Bearing this in mind, a seed was considered “detected” in this demonstration if it was within approximately 0.60 m of its closest anomaly, close enough to be recovered in a real remediation project. As shown in the histogram, all 160 seeds were within 0.60 m of their closest EM61-Mk2 anomaly, indicating that all seeds were detected by the EM61-Mk2 cart. In fact, the maximum distance between a seed and its closest anomaly was only 0.46 m, well below the 0.60 m distance threshold. The Pd of the EM61-Mk2 cart was therefore estimated as Pd = 160 / 160 = 100%. The 95% confidence interval around this Pd was estimated as 98%–100%, based on the exact binomial distribution. That is to say, if this demonstration could be repeated at 100 different sites exactly like the former Camp Butner, then 95 times out of 100, the EM61-Mk2 cart would exhibit a Pd between 98% and 100%, with respect to the seeded UXO.

The FAR was also estimated for the EM61-Mk2 cart. A total of 2304 anomalies had been listed on the EM61-Mk2 anomaly list. Of these 2304 anomalies, 160 detected the seeded UXO, while the remaining 2144 did not. Had this been a real remediation project, UXO technicians would have attempted to recover a target from these 2144 remaining anomalies. It is possible that some native UXO would have been recovered. At this point in the process, though, the locations of native UXO were not yet known. As a worst-case estimate, it was assumed that there were *no* native UXO, such that all remaining anomalies were false alarms. Therefore, FAR was estimated as the number of remaining anomalies divided by the acreage of the demonstration area: FAR = 2144 / 4.4 acres = 487/acre.

Table 3: Distances between seeds and their closest EM61-Mk2 and MetalMapper anomalies.

Summary Statistic	EM61-Mk2 Cart	MetalMapper
N	160	160
Minimum (m)	0.02	0.02
Maximum (m)	0.46	3.08
Mean (m)	0.20	0.20
Standard deviation (m)	0.10	0.27
Median (m)	0.19	0.16
Inter-quartile range (m)	0.13	0.14

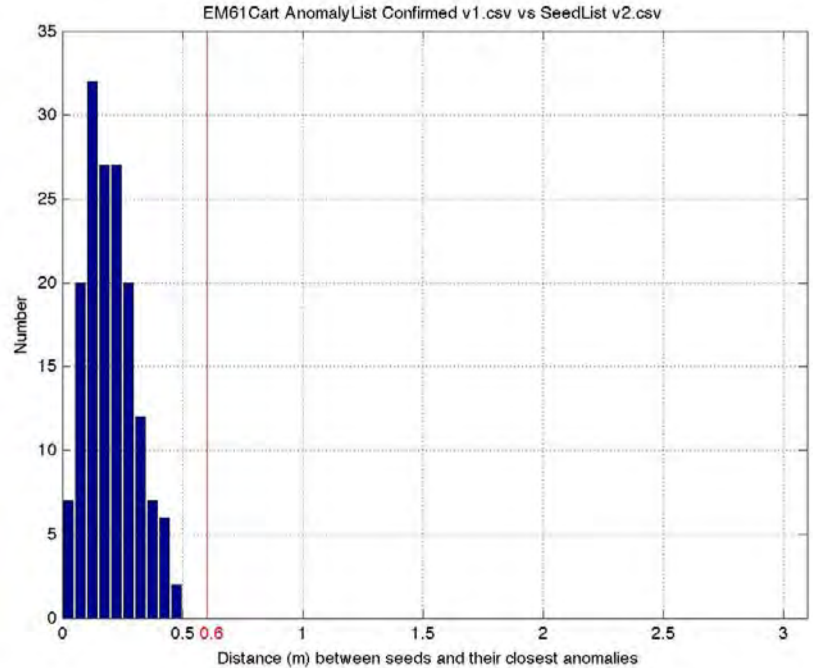


Figure 13: Distances between seeds and their closest EM61-Mk2 anomalies. All seeds were closer than 0.60 m (2 ft) to their closest anomaly.

A two-step process was also used to estimate Pd for the MetalMapper. Figure 14 is a histogram of the distances between the seeds and their closest MetalMapper anomalies, with these distances summarized in Table 3. Three seeds were flagged as being farther than 0.60 m from their closest MetalMapper anomaly. ESTCP closely investigated the MetalMapper dynamic data collected around each of these three seeds to determine if these seeds could be considered “detected” (i.e., if these seeds would have been recovered in a real remediation project).

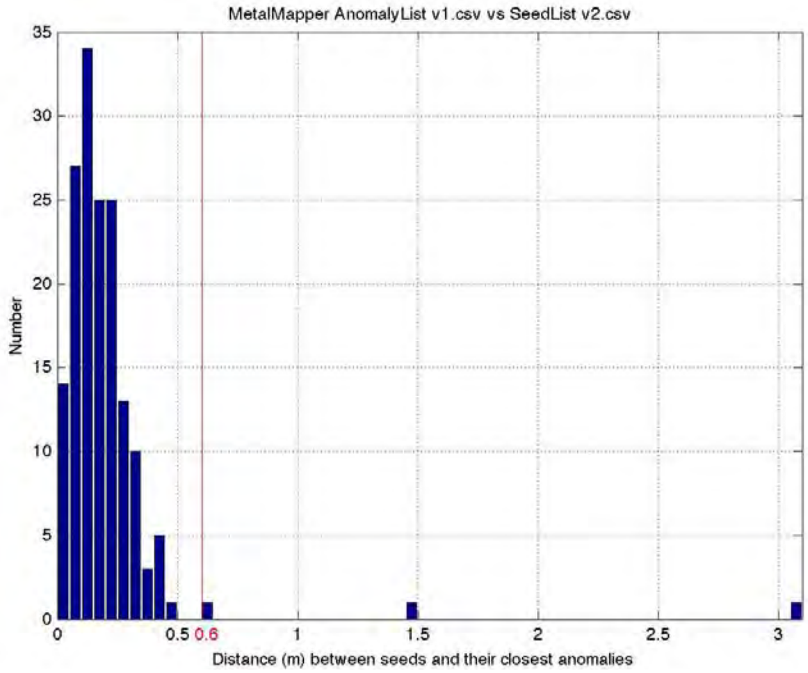


Figure 14: Distances between seeds and their closest MetalMapper anomalies. Three seeds were farther than 0.60 m (2 ft) from their closest anomaly.

Further investigation showed that the MetalMapper did in fact detect seed #152, the first of the three seeds farther than 0.60 m from its closest MetalMapper anomaly. This seed was a 37 mm projectile buried at 30 cm. Figure 15(a) shows a photograph of the seed immediately before burial, and Figure 15(b) shows a 10 m × 10 m grid of the MetalMapper dynamic data collected around the seed after burial. A purple circle of radius 0.30 m (1 ft) indicates the seed location. Black circles, also with radii of 0.30 m, indicate the locations of the detected MetalMapper anomalies. By definition, all detected anomalies had amplitudes greater than 2.0 mV, the final MetalMapper detection threshold. As shown in the figure, seed #152 was only 0.62 m from its closest MetalMapper anomaly. Had this been a real remediation project, UXO technicians would have been able to recover this seed with the use of a hand-held instrument.

The MetalMapper did not detect seeds #101 and #104, the two other seeds farther than 0.60 m from their closest MetalMapper anomalies. Both seeds were 37 mm projectiles buried at 30 cm. Figure 15(c) and (e) show photographs of the seeds immediately before burial, and Figure 15(d) and (f) show 10 m × 10 m grids of the MetalMapper dynamic data surrounding the seeds after burial. In both cases, the seeds gave rise to weak anomalies whose amplitudes did not surpass the MetalMapper detection threshold, as is indicated by the small gray blobs directly underneath the purple seed circles in the figures. Had this been a real remediation project, these seeds would not have been recovered. Thus the Pd of the MetalMapper was estimated as: Pd = 158 / 160 = 99%. The 95% confidence interval was calculated as 96%–100%.

The FAR of the MetalMapper was calculated as follows. Of the 3765 anomalies on the final MetalMapper anomaly list, 158 of them had detected a seeded UXO. Assuming that the remaining 3607 anomalies did not detect native UXO, then FAR = 3607 / 4.4 acres = 819/acre.

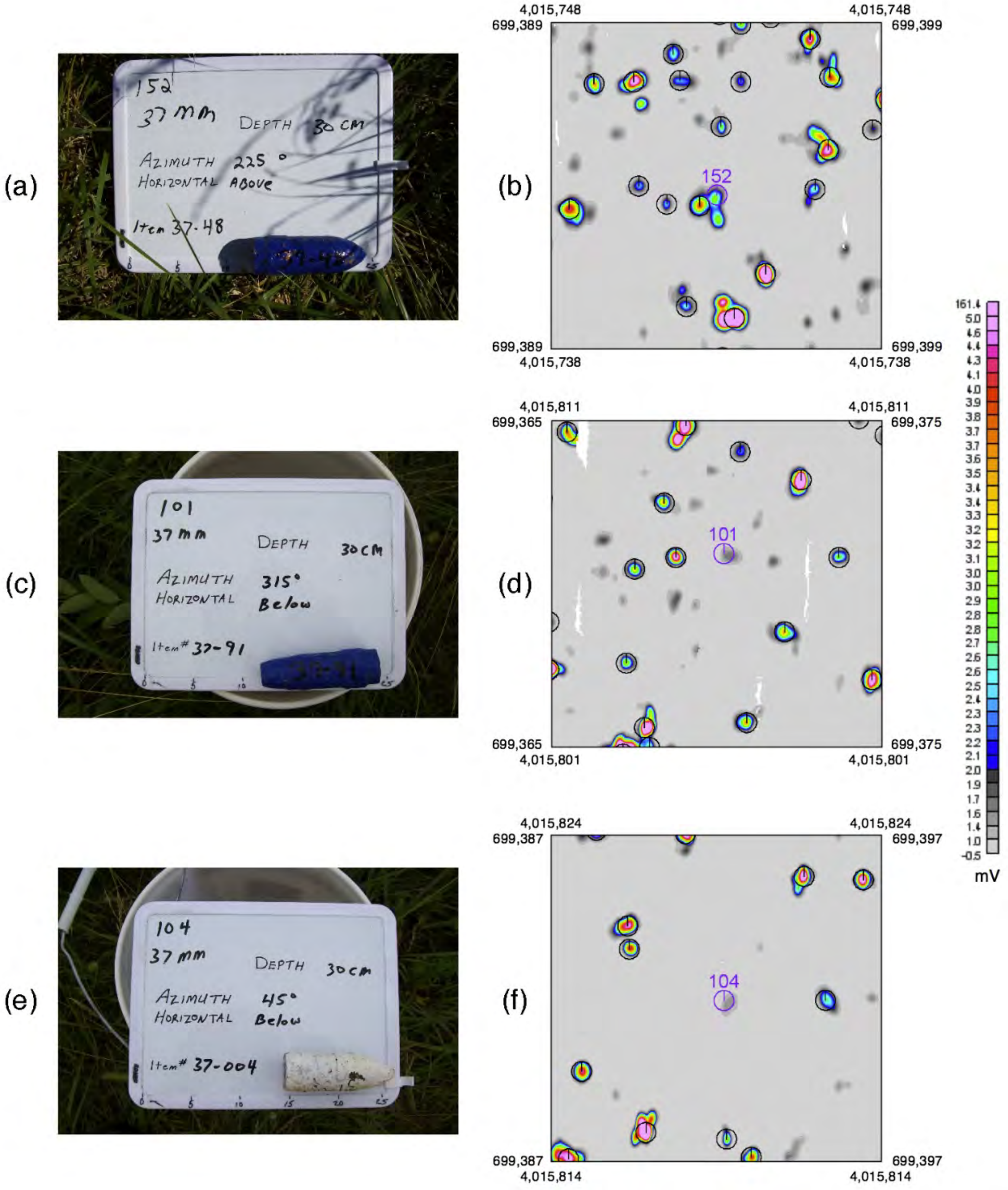


Figure 15: (a, c, e) Photographs of three seeds. (b, d, f) 10 m × 10 m grids of MetalMapper dynamic data surrounding the seeds. All three seeds were farther than 0.60 m (2 ft) from their closest MetalMapper anomaly. Purple circles indicate the seed locations while black circles mark the anomaly locations. All circles have radii of 0.30 m (1 ft). All detected anomalies (black circles) had amplitudes greater than the final MetalMapper detection threshold of 2.0 mV.

In summary, the EM61-Mk2 cart exhibited a better detection performance than the MetalMapper, based on the seeded UXO. The cart detected even the smallest seeds (37 mm projectiles) buried at the greatest depths (30 cm), leading to a Pd of 100% (98%–100%) and a FAR of 487/acre. In contrast, the MetalMapper failed to detect two 37 mm projectiles at 30 cm depths, leading to a Pd of 99% (96%–100%) with a FAR of 819/acre, almost twice that of the EM61-Mk2 cart. The difference in Pd between the two instruments was not statistically significant, as there was overlap between the two 95% confidence intervals. However, the Advisory Group commented that had this been a real remediation project, stakeholders would have been troubled by the MetalMapper's inability to detect two seeds. These two seeds gave rise to weak anomalies that did not surpass the MetalMapper detection threshold, even though a 50% safety margin had been used to lower the threshold from its original level, resulting in the high FAR. MetalMapper used a wider line spacing than the EM61-Mk2 cart (0.75 m versus 0.50 m), and one of the triaxial receive coils on one of the MetalMapper instruments functioned only intermittently at best [47][50][53]. It is possible that these factors resulted in the MetalMapper's poorer detection performance.

D. Collecting Static Data

Once anomalies were detected in the dynamic data, two types of EMI instruments were used to collect static data: the TEMTADS and two MetalMapper systems (the same MetalMappers that were used to collect dynamic data). Brief descriptions of the instruments and data-collection methods are given below, with more detail available in the documents written by the data-collection teams [3][53][54][60][61].

1. TEMTADS

The TEMTADS is an advanced instrument developed by NRL [3][61]. Its design is based on the Advanced Ordnance Locator system, developed by G&G Sciences under Navy funding. As shown in Figure 16, the TEMTADS employs 25 sensors arranged in a 5 × 5 array. Each sensor consists of a 35 cm square outer transmit coil and a 25 cm square inner receive coil. The coils are mounted with their centers 40 cm apart, producing a 2 m × 2 m square array. Each transmit coil is pulsed in sequence, and the secondary field induced by a buried target is sensed simultaneously by all 25 receive coils. As the TEMTADS is deployed in static mode, it rests at any given location for several seconds while successive measurements are "stacked," or averaged, over time. Measurements are sampled at a rate of 500 kHz and then grouped into 115 time gates ranging from 42 μs to 25 ms. The TEMTADS also employs three GPS antennas to determine the location and orientation of the sensor array.

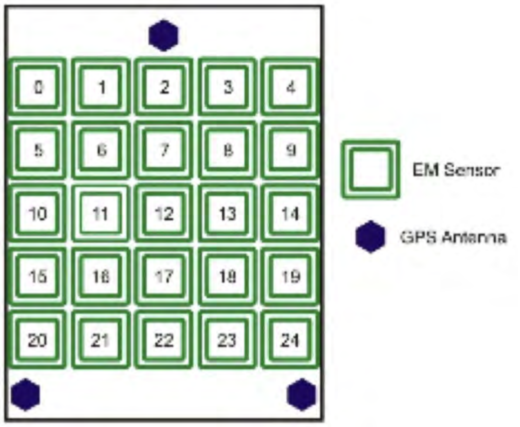


Figure 16: The TEMTADS consists of 25 sensors arranged in a 5 × 5 array. Each sensor consists of an outer transmit coil and an inner receive coil. Each transmit coil is pulsed in sequence to produce a primary electromagnetic field, after which all receive coils simultaneously sense the secondary electromagnetic field induced in the buried target. Taken from [29].

Nova Research Inc. collected static data with the TEMTADS at the former Camp Butner, as shown in Figure 17 [3][61]. The instrument was calibrated at the IVS at the beginning and end of each day. The training pit was used to collect data from the expected UXO at the different depths and orientations requested by the classification analyst teams. In the demonstration area, Nova drove the TEMTADS to a location on the EM61-Mk2 anomaly list, visually checked contour plots of the sensed signal to make slight adjustments to the position of the sensor such that it was more likely to be located directly above the buried target, acquired static data from the buried target, and then moved on to the next location on the list. Later, the data-collection team “inverted,” or analyzed, each static data set to characterize the buried target. If this process failed for any reason (e.g., the SNR was too low, the sensor had been incorrectly positioned, etc.), then the data-collection team returned to that location and reacquired static data. Thus, each EM61-Mk2 anomaly corresponded to one or more TEMTADS static data sets.

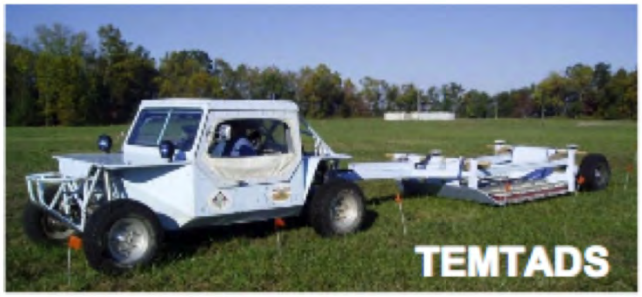


Figure 17: Nova Research Inc. operated the TEMTADS at the former Camp Butner. Taken from [51].

There are advantages and disadvantages of using the TEMTADS for UXO classification. One advantage is that the 5 × 5 sensor array provides spatial diversity for both transmitting and receiving, allowing the primary field to illuminate the buried target in all directions and allowing the secondary field to be sensed in all directions. Furthermore, because the coils are mounted a fixed distance from each other, their relative position accuracy is very high. In addition, stacking boosts the SNR of the signal. Finally, the time gates for the TEMTADS stretch out to 25 ms, much longer than either of the instruments used in dynamic mode, allowing a longer temporal extent for assessing the decay of the secondary field over time. All of this results in a much more accurate characterization of the buried target, in comparison to either of the two dynamic instruments. However, the TEMTADS does have some limitations. As a static sensor, it is slower to deploy; the operator is required to park the instrument over each location for several seconds at a time. Also, the operator must be told in advance at which locations to collect data; another instrument operating in dynamic mode is required to detect the anomalies in the first place. Finally, the TEMTADS is a new instrument that has not been transferred to the commercial pipeline. Few individuals know how to operate the instrument or analyze its data.

2. MetalMapper

The MetalMapper can be used in dynamic or static mode. In static mode, each of the three orthogonal transmit coils are excited in sequence, illuminating the buried target from three orthogonal directions, one at a time. (In dynamic mode, only one transmit coil is used.) As in dynamic mode, the strength of the secondary field is measured simultaneously in each of the seven triaxial receive coils. The latest time gate is generally set much later in static mode than in dynamic mode, as the MetalMapper can afford to wait several seconds for data collection when deployed in static mode [53].

Geometrics Inc. operated the MetalMapper in static mode at the former Camp Butner. Its subcontractor, Sky Research Inc., simultaneously collected static data using a second, newer system [53][54][60]. Both organizations configured their time gates from 106 μs to 7.9 ms. The instruments were calibrated daily at the IVS. Static data were collected in the training pit, as had been done by the other instruments. In the demonstration area, the team attempted to collect one static data set for each anomaly detected in the MetalMapper dynamic data, as well as one static data set for each EM61-Mk2 anomaly farther than 0.60 m from any MetalMapper anomaly. The team experienced many difficulties in the field, however, leading to questions about which static data sets corresponded to which MetalMapper anomalies and which EM61-Mk2 anomalies. These ambiguities could have caused significant problems in the eventual scoring of the classification analyses. Therefore, Geometrics identified which of its static data sets corresponded to which EM61-Mk2 anomalies, such that all static data analyses could be tied to the EM61-Mk2 anomaly list only, as opposed to both the EM61-Mk2 and MetalMapper lists. In the end, all EM61-Mk2 anomalies corresponded to one or more MetalMapper static data sets.

There are advantages and disadvantages to using the MetalMapper in static mode for UXO classification. As with the TEMTADS, MetalMapper's multiple coils provide spatial diversity for transmitting and receiving. The relative position accuracy between coils is very high because the coils are mounted at fixed distances with respect to each other. Stacking boosts the SNR of the signal. All of this leads to an excellent characterization of the buried target. In addition, although the MetalMapper is a new instrument, commercial sales have already begun, and more and more individuals are becoming familiar with its operation. The MetalMapper in static mode does have some limitations, however. In this study, its longest time gate (7.9 ms) occurred sooner than that of the TEMTADS (25 ms), making it more difficult to assess the secondary field's decay over time. In addition, like the TEMTADS, the MetalMapper in static mode is slower to operate than either the EM61-Mk2 cart or itself in dynamic mode. Finally, the MetalMapper in static mode requires a list of locations at which to collect static data (although as a dual-mode instrument, it could help provide this information by first collecting data in dynamic mode).

4. Collecting Ground Truth

Scientific demonstrations often require ground truth. In a live-site UXO classification demonstration, all targets that gave rise to a detected anomaly must be recovered and catalogued. A fraction of the catalogued information can be used as training data to optimize the algorithms that classify the detected anomalies. The remainder of the catalogued information can be used as ground truth to score the classification analyses. This section describes the methods used at the former Camp Butner to create a list of locations from which to recover buried targets, the methods used to recover the targets, and the philosophy used to assign ground-truth labels to the detected anomalies based on their recovered targets.

A. Creating the Recovery List

The recovery team required a list of easting and northing coordinates from which to recover buried targets. These targets were meant to be those that gave rise to all EM61-Mk2 anomalies, since these were the only anomalies considered in the classification analyses. Therefore, ESTCP briefly considered guiding the recovery with the list of EM61-Mk2 anomaly locations (the same list that was used to guide the collection of static data). However, the EM61-Mk2 anomalies were merely the peaks in the gridded dynamic data and were not necessarily located directly above the buried targets. As was discussed in detection scoring, recovery teams often use a GPS unit to reacquire the location of a detected anomaly and then use a hand-held magnetometer or EMI instrument to probe the ground within 0.60 m (2 ft) to pinpoint the exact location of the buried target. Indeed, Table 3 showed that the EM61-Mk2 anomalies were a mean distance of 0.20 m from the seeded UXO. Furthermore, the anomalies at the former Camp Butner were quite dense, with many anomalies only slightly farther than 0.60 m from each other. Using the EM61-Mk2 anomaly list to guide the recovery could therefore have led to some ambiguities in recovering the buried targets, since holes could have been dug in a slightly different location than was intended. This could have resulted in a nearby target mistakenly recovered in place of the target that had produced the anomaly in the first place.

Other lists were used to guide the recovery of buried targets. The TEMTADS and MetalMapper data-collection teams had already “inverted,” or processed, each static data set to characterize the buried target’s size, shape, material composition, and wall thickness, as well as to provide a better estimate of the target’s location. IDA assessed the position accuracy of the TEMTADS and MetalMapper static inversion locations by comparing them to the locations of the seeded UXO. The distance between each seed and its closest static inversion was measured, similar to the process used to score the detection performance of the dynamic instruments (summarized in Table 3, Figure 13, and Figure 14). Table 4 summarizes these measured distances, and Figure 18 and Figure 19 show histograms of them. The inversion locations from both static instruments displayed better position accuracy than the anomaly locations from either dynamic instrument, exhibiting tighter histograms and shorter mean and median distances. This result was expected for two reasons: (1) When collecting static data, both data-collection teams inspected real-time data (e.g., contour plots, etc.) to better position the sensor directly over the buried target. (2) Data inversions are known to produce better position estimates of the buried target than the raw data themselves. The TEMTADS exhibited the best position accuracy, likely due to the careful field technique exhibited by its data-collection team.

Table 4: Distances between seeds and their closest TEMTADS and MetalMapper static inversions.

Summary Statistic	TEMTADS	MetalMapper
N	160	160
Minimum (m)	0.00	0.02
Maximum (m)	0.45	0.43
Mean (m)	0.05	0.10
Standard Deviation (m)	0.05	0.07
Median (m)	0.03	0.07
Inter-Quartile Range (m)	0.03	0.10

Based on these results, ESTCP relied heavily on the TEMTADS static inversion locations to create the recovery list. The EM61-Mk2 dynamic data map from Figure 11 was overlaid with the locations of the EM61-Mk2 anomalies, as well as the TEMTADS and MetalMapper static inversion locations. For each anomaly, ESTCP visually inspected the surrounding EM61-Mk2 data and subjectively determined which specific easting and northing coordinates should be entered on the recovery list. For most of the 2304 anomalies, the location of the closest TEMTADS static inversion was used. In a few hundred cases, though, it appeared that the TEMTADS had not succeeded in collecting static data directly above the target. In these cases, the EM61-Mk2 anomaly locations were entered on the recovery list instead. In only one case was the MetalMapper static inversion location used. Finally, in 120 cases, the TEMTADS static inversion location was farther than 0.60 m from the EM61-Mk2 anomaly location. Both locations were included on the recovery list, in case two targets were actually buried in the ground.

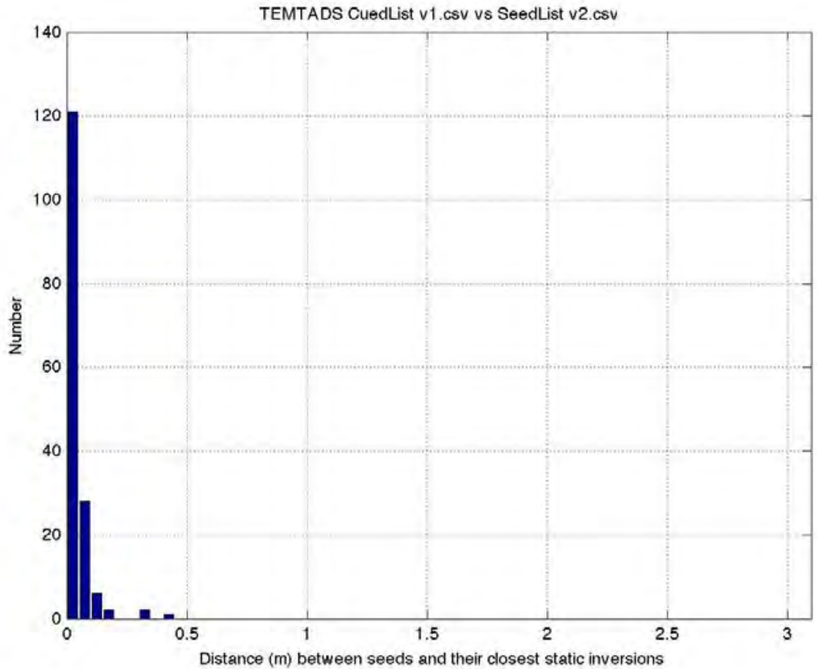


Figure 18: Distances between seeds and their closest TEMTADS static inversions.

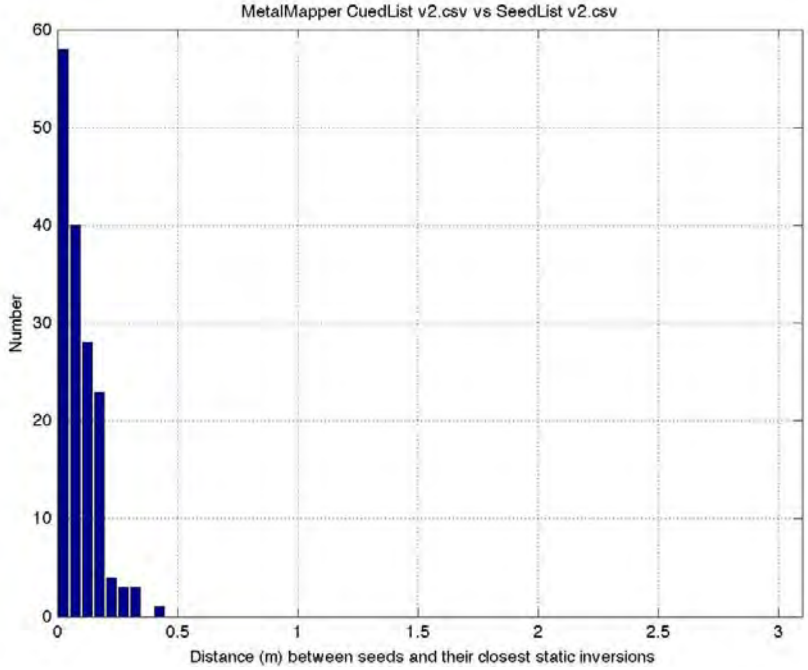


Figure 19: Distances between seeds and their closest MetalMapper static inversions.

B. Recovering Buried Targets

Parson Inc. recovered all targets buried at each location on the recovery list. In most cases, a single target was recovered from each location. In some cases, though, multiple targets were recovered from the same location, and other locations had no targets. (In situations where no target was found, ESTCP instructed Parsons to use a hand-held instrument to probe the inside of the dug hole, and to cease digging only if no anomaly was detected.) Upon uncovering a target, but before removing it from the ground, Parsons catalogued the following information:

- The easting, northing, and depth coordinates of the center of the target, with depth measured with respect to the average of one or more surveyed points on the lip of the hole.
- A description of the target (e.g., “105 mm HEAT,” “Frag,” “Metal Plow,” etc.).
- A photograph of the target alongside a ruler and a whiteboard listing the Target ID. Figure 20 shows photographs of six different recovered targets.

C. Assigning Ground Truth Labels to the Recovered Targets

ESTCP assigned a single ground-truth label to each anomaly based on its recovered targets. Specifically, an anomaly was labeled “TOI” if any of its recovered targets was a seeded or native UXO of any type, such as those shown in Figure 20(a)–(e). Conversely, an anomaly was labeled as “Non-TOI” if none of its recovered targets were seeded or native UXO, such as that shown in Figure 20(f).

All 160 seeded UXO were recovered, along with 11 native 37 mm projectiles found at depths of 2–18 cm. Seven of the native UXO posed a risk of detonation. The remaining four were empty shells with hazardous material already spent. Although these four could not have exploded, ESTCP deemed them TOIs because in a real-world situation, they would appear to be hazardous, and safety-oriented procedures would be used for their recovery.

Because each UXO was recovered from a different anomaly, 171 of the 2304 anomalies were labeled “TOI.” Another 2121 anomalies were labeled as “Non-TOI.” Most of the recovered Non-TOIs were munitions debris, but some were cultural debris (e.g., wrenches, pieces of fencing, etc.) and others were labeled “No contact” (i.e., no targets were found for recovery, indicating that the anomalies had been produced by either geology or noise).

No attempts were made to recover targets from 12 of the 2304 anomalies. Two of these anomalies had been inadvertently left off of the recovery list. Due to lack of ground-truth information, these two anomalies were removed from all further analyses. The final 10 anomalies had been purposely left off the recovery list, since visual inspection of the EM61-Mk2 data maps had indicated that they had likely been produced by the same buried targets as neighboring anomalies. Therefore, these 10 anomalies were also removed from all further analyses. In the end, classification was attempted for only those 2292 anomalies for which ground truth was known. Table 5 gives the details.

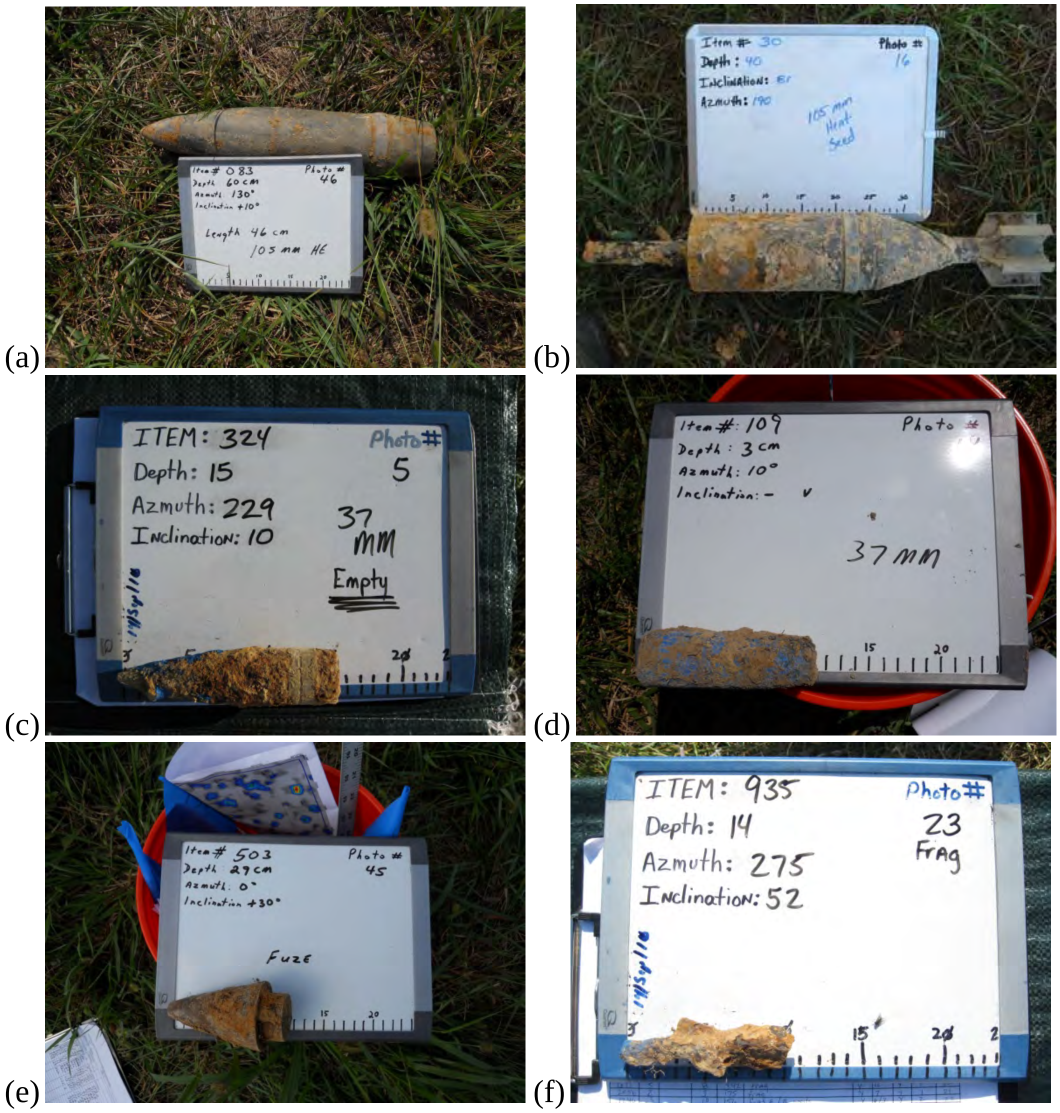


Figure 20: Photographs of recovered targets: (a) 105 mm projectile, (b) 105 mm HEAT projectile, (c) 37 mm projectile with driving band, (d) 37 mm projectile without driving band, (e) M48 fuze, and (f) scrap metal from a previously exploded munition. The targets in (a)–(e) were labeled "TOI"; the target in (f) was labeled "Non-TOI."

Table 5: Ground truth at the former Camp Butner: 2292 of the 2304 anomalies detected in the EM61-Mk2 data were excavated. TOIs were recovered from 171 anomalies, including the 160 seeded UXO plus 11 native 37 mm projectiles. No TOIs were recovered from 2121 anomalies. The remaining 12 anomalies were not excavated.

Seeded TOIs		
Type	**Number**	**Depth**
37 mm projectile	110	3–35 cm
105 mm projectile	26	10–62 cm
M48 fuze	24	7–33 cm
Native TOIs		
Type	**Number**	**Depth**
37 mm projectile	11	2–18 cm
Native Non-TOIs		
Type	**Number**	**Depth**
Munitions debris	2041	0–155 cm
Cultural debris	41	0–20 cm
No contact	39	Not applicable

5. Classifying Data

UXO classification is a multistep process. First, the collected data are analyzed to estimate the characteristics of the buried target (e.g., features related to the size, shape, material composition, and wall thickness of the buried target). Next, the estimated characteristics are used to classify the buried targets as either likely UXO or likely clutter. Computer-based algorithms are used to perform both the target characterizations and classifications.

Six different classification analyst teams participated in the demonstration: CH2M HILL, Dartmouth College, Geometrics Inc., NAEVA Geophysics Inc., SAIC, and Sky Research Inc. NAEVA and SAIC also performed retrospective analyses after the demonstration was completed and full ground truth was released to the public. Two additional organizations, Parsons Inc. and Signals Innovations Group (SIG), performed retrospective analyses only. CH2M HILL, NAEVA, and Parsons are commercial geophysics companies that often perform UXO remediation in real-world situations. Dartmouth, Geometrics, SAIC, and SIG are organizations involved in the research and development of advanced UXO classification technologies. Sky is involved in real-world remediation projects, as well as research and development efforts. As shown in Table 6, each team used different methods for classifying the detected anomalies. This section gives a brief description of these methods. More detail can be found in the documents written by the classification analyst teams [35][39][42][46][47][50][54][56].

Table 6: Classification analyst teams at the former Camp Butner. Retrospective analyses are shaded in gray.

Classification Analyst Team	EM61-Mk2 cart dynamic data only	All TEMTADS static data	EM61-Mk2 cart dynamic data + TEMTADS static data requests	All MetalMapper static data	EM61-Mk2 cart dynamic data + MetalMapper static data requests	MetalMapper dynamic data only
CH2M HILL				**Features:** Polarizabilities from UXOLab dipole model **Algorithm:** Library match with and without human expert **Training:** Custom		
Dartmouth		**Features:** Full curves from ONVMS non-dipole model **Algorithm:** Library match **Training:** Custom		**Features:** Full curves from ONVMS non-dipole model **Algorithm:** Statistical classifier **Training:** Custom		
Geometrics				**Features:** Amplitudes, ratios, and decays of polarizabilities from MMRMP dipole model **Algorithm:** Rules, library match, and neural network **Training:** Standard		

Classification Analyst Team	EM61-Mk2 cart dynamic data only	All TEMTADS static data	EM61-Mk2 cart dynamic data + TEMTADS static data requests	All MetalMapper static data	EM61-Mk2 cart dynamic data + MetalMapper static data requests	MetalMapper dynamic data only
NAEVA	**Features:** Amplitudes, footprints, and decay rates of data with UXDetect and UXProcess **Algorithm:** Rules **Training:** Standard			**Features:** Polarizabilities from UXAnalyze dipole model **Algorithm:** Library match **Training:** Standard		
Parsons	**Features:** Decays of polarizabilities from custom and UXAnalyze dipole models **Algorithm:** Rules **Training:** Custom				**Features:** Amplitudes, ratios, and decays of polarizabilities from UXAnalyze dipole model **Algorithm:** Library match **Training:** Custom	
SAIC	**Features:** Sum and decays of polarizabilities from UXAnalyze dipole model **Algorithm:** Statistical classifier **Training:** Standard	**Features:** Amplitudes, ratios, and decays of polarizabilities at all time gates from custom and UXAnalyze dipole models **Algorithm:** Library match **Training:** None and Custom	**Features:** Amplitudes, ratios, and decays of polarizabilities from UXAnalyze dipole model **Algorithm:** Library match **Training:** None	**Features:** Amplitudes, ratios, and decays of polarizabilities at all time gates from custom dipole model **Algorithm:** Library match **Training:** None and Custom	**Features:** Amplitudes, ratios, and decays of polarizabilities from UXAnalyze dipole model **Algorithm:** Library match **Training:** Custom	

Classification Analyst Team	EM61-Mk2 cart dynamic data only	All TEMTADS static data	EM61-Mk2 cart dynamic data + TEMTADS static data requests	All MetalMapper static data	EM61-Mk2 cart dynamic data + MetalMapper static data requests	MetalMapper dynamic data only
Sky	**Features:** Sum and decays of polarizabilities from UXOLab dipole model **Algorithm:** Statistical classifier **Training:** None	**Features:** Full polarizability curves from UXOLab dipole model **Algorithm:** Library match and statistical classifier **Training:** Custom		**Features:** Full polarizability curves from UXOLab dipole model **Algorithm:** Library match and statistical classifier **Training:** Custom	**Features:** Full polarizability curves from UXOLab dipole model **Algorithm:** Statistical classifier **Training:** Custom	**Features:** Sum and decay of polarizabilities from UXOLab dipole model **Algorithm:** Library match with and without human expert and statistical classifier **Training:** Custom
SIG		**Features:** Amplitudes of polarizabilities at two time gates from custom dipole model **Algorithm:** Statistical classifier **Training:** Custom				

A. Requesting Static Data

Classification can require different types of data, some of which can be expensive to collect. While dynamic data can be used to classify the detected anomalies under benign conditions, the higher quality static data are often required when target types become more challenging. Collecting high-quality static data is expensive because advanced instruments are expensive to purchase or lease and their operation is time consuming. Therefore, classification methods that rely on only a small amount of static data could be advantageous to the UXO community.

Each classification analyst team determined for itself which anomalies required static data. Different teams used different criteria for making this determination. The teams submitted their static data requests to ESTCP, and ESTCP distributed the static data to the teams only upon request. Some teams, such as CH2M HILL, Dartmouth, and Geometrics, requested static data for all anomalies, regardless of the quality of the dynamic data [28][42][52]. (SIG also requested static data for all anomalies in its retrospective analysis, once the demonstration was complete and ground truth had been released to the public [35].) NAEVA did the opposite; this team did not request static data for any anomalies and analyzed only the EM61-Mk2 dynamic data [18]. (NAEVA also performed retrospective analyses of the MetalMapper static data for all anomalies [47].) SAIC and Sky did both; in one set of analyses, they analyzed only the EM61-Mk2 dynamic data; in another set, they analyzed only the TEMTADS or MetalMapper static data [1][22]. In addition, these two teams also performed a third type of analysis, in which they first analyzed the quality of the EM61-Mk2 dynamic data and then requested static data for only a subset of the anomalies for which the dynamic data was not adequate [1][22]. (Likewise, Parsons performed some retrospective analyses using both the EM61-Mk2 dynamic data and some requested MetalMapper static data [44][45]. They also performed other retrospective analyses using only the EM61-Mk2 dynamic data [46].) Sky also performed a fourth type of analysis, in which they analyzed only the MetalMapper dynamic data [22].

Teams with more than one type of analysis made efforts to keep their analyses “blind.” To do this, they implemented internal firewalls between their individual team members. For example, SAIC reported that the analyst processing the TEMTADS data did not share any data, results, or insights with any other analysts processing the EM61-Mk2 or MetalMapper data [1]. In this way, information gleaned from one data set did not bias other analyses using other data sets.

B. Requesting Ground Truth

In a real remediation project, the purpose of ground truth during classification is to “train,” or optimize, the parameters on which the classification algorithms are based.

However, collecting ground truth is expensive because time-consuming, safety-oriented precautions must be followed when recovering targets, since any one of them could turn out to be UXO. As such, the UXO community desires classification methods that require little to no new ground truth for training. Bearing this in mind, this demonstration was designed such that the classification analyst teams could make their own decisions regarding what ground truth was needed for training.

The classification analyst teams were allowed to train their algorithms using little to no ground truth collected from the former Camp Butner. Instead, many teams used ground truth collected in the previous two demonstrations at the former Camp Sibert and former Camp San Luis Obispo, as well as other previous studies at Aberdeen and Yuma Proving Grounds. In addition, the teams were also given the option to use the limited ground truth collected from the IVS and training pit at the former Camp Butner. The teams could input the data into their classification algorithms and then automatically and systematically adjust the parameters of the algorithms to give results as close as possible to the known ground truth. Some teams, such as SAIC and Sky, were able to optimize their algorithms using ground truth from only the IVS, training pit, and previous studies [1][22]. Other teams required more extensive information from the Butner demonstration area itself.

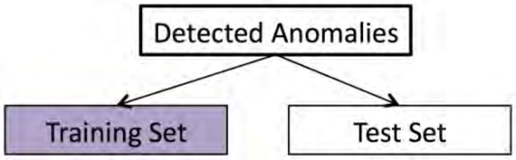
IDA compiled a “Standard Training Set” for more site-specific training. The set was composed of all anomalies in a 30 m × 30 m grid. There were 179 anomalies in this grid, 173 of which had been assigned the ground-truth label of Non-TOI. Only six had been labeled TOI. Four of the TOIs were 37 mm projectiles, and two were M48 fuzes. IDA chose this section of the site for the Standard Training Set because it was located in an obvious section of the demonstration area (one of the two most southern grids) and was suitably dense to provide a sufficient number of anomalies that were adequately spaced from each other. Two classification analyst teams, Geometrics and NAEVA, chose to use the Standard Training Set to train their classification algorithms [18][52]. SAIC also used the Standard Training Set in its retrospective analysis.

Other teams chose to compile their own custom-built training sets. Each of these teams assessed the data collected for each anomaly in the Butner demonstration area and determined which anomalies could best optimize their classification algorithms. They submitted their requests to ESTCP, and ESTCP distributed the ground truth (identities, photographs, locations, depths, and orientations) upon request. Some requests were made in series; a team first submitted a request for only a handful of anomalies, optimized their classification algorithms based on this set, decided that additional ground truth was needed, and then submitted a second (and third and fourth, etc.) request for additional anomalies. CH2M HILL and Dartmouth requested custom training sets for all of their classification analyses, as did Parsons and SIG in their retrospective analyses [35][42][44][45][52]. SAIC and Sky performed several different types of analyses.

Although some of these analyses used algorithms optimized over only the ground truth collected in the IVS, training pit, and previous studies, other analyses used algorithms optimized over a custom training set consisting of a subset of the anomalies from the Butner demonstration area [1][22].

Teams with more than one type of training set made efforts to keep their analyses blind. For example, SAIC reported that the analyst using an algorithm optimized over only the IVS, training pit, and previous studies first submitted his original classification deliverable to ESTCP. Only then did he request a custom training set consisting of anomalies from the Butner demonstration area. The analyst used the custom training set to re-optimize his classification algorithm and create a final classification deliverable [1].

All anomalies that were *not* assigned to the training set were assigned to the complementary test set, as illustrated in Figure 21. For example, both Geometrics and NAEVA used the Standard Training Set to optimize their classification algorithms. All other anomalies in the demonstration area were therefore assigned to the test set. In another example, one of Sky's analyses used only information from the IVS, training pit, and previous studies for algorithm optimization. *All* anomalies from the demonstration area, then, were assigned to the test set. In a final example, each of Sky's remaining analyses used a custom training set consisting of some anomalies from the demonstration area. All anomalies in the demonstration area *not* assigned to the custom training set were assigned to the complementary custom test set. Regardless of how the anomalies were assigned to the training or test sets, each team was required to classify each anomaly in the test set. The following sections describe how this was accomplished.


Detected Anomalies

Training Set

Test Set

Figure 21: Each classification analyst team separated the EM61-Mk2 anomalies into complementary training and test sets.

C. Defining "Cannot Analyze"

The classification analyst teams separated the test set anomalies into two groups, "Can Analyze" and "Cannot Analyze," as shown in Figure 22. This was done based on the quality of the collected data (both the dynamic data and, where available, the static data). For most anomalies, the collected data were of sufficient quality to accurately characterize the buried targets. These anomalies were put into the "Can Analyze" group. For some anomalies, however, the collected data suffered from geolocation errors, spotty coverage, low data density, or low SNR, making it difficult, if not impossible, to accurately characterize the buried targets. These anomalies were put into the “Cannot Analyze” group.

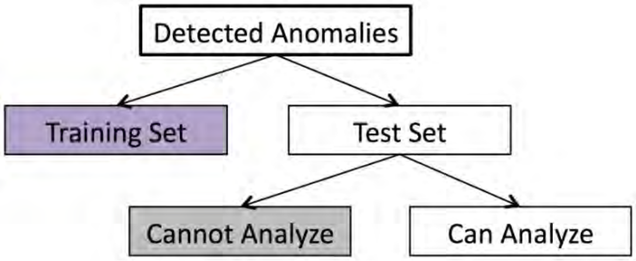


Figure 22: Each classification analyst team further separated the test set anomalies into “Can Analyze” and “Cannot Analyze” groups.

Different classification analyst teams used different criteria to separate the “Can Analyze” and “Cannot Analyze” anomalies. Some teams, such as Dartmouth, Geometrics, NAEVA, and SAIC, used quantitative criteria, such as the match score of how well the collected data fit to a geophysical model [6] [10] [11] [12] [13] [14] [15] [16] [22] [55]. Other teams, such as Parsons, used subjective criteria, such as visual analysis of the collected data [44][45][46]. Furthermore, in many cases, the same team put an anomaly into the “Can Analyze” group based on one instrument’s data, but into the “Cannot Analyze” group based on another instrument’s data because different instruments have different resolutions, SNR, etc.

D. Characterizing the Buried Targets

The classification analyst teams estimated the characteristics of the buried targets that produced the “Can Analyze” anomalies. Most teams estimated physical properties of the targets based on a geophysical model of the collected data. In contrast, one team made direct measurements of the data itself. In either case, each team then selected features on which classification would be based.

1. Model-driven Characterizations

Five teams—CH2M HILL, Dartmouth, Geometrics, SAIC, and Sky [3][22][28][42][46]—used geophysical models to process the available data. (NAEVA, Parsons, and SIG also used geophysical models in their retrospective analyses [6][35][44][45].) A geophysical model is a set of equations that estimate the data that would be produced by a target of known size, shape, material composition, and wall thickness buried at a known depth and orientation. Many models assume that the target can be represented as one or more point dipole sources, an accurate assumption when the target is isolated from other targets and when the target is not within the near field of the sensor. More advanced models recently introduced to the UXO community do not assume dipole sources.

Geophysical models can be used to estimate the characteristics of a buried target. To do this, the target characteristics are first assumed to coincide with a set of initial conditions, and the model estimates the data that would be produced by such a target. Changing any one of these conditions can lead to a change in the modeled data. This is called the “forward model.” The analyst then inputs the actual data collected for an anomaly into a computer-based algorithm. The algorithm compares the collected data to the modeled data and calculates a match score. Using an optimization procedure, the algorithm adjusts one or more of the modeled target’s characteristics, re-estimates the data that the modeled target would produce, and compares the re-modeled data to the collected data. The process iterates until the match score is optimized. At that point, the characteristics of the buried target are estimated to be the characteristics of the most recently modeled target. This process is known as “inversion” because it uses at the forward model in a reverse, or “inverse,” manner. For example, Figure 23(a) shows the EM61-Mk2 data collected for one anomaly at the former Camp Butner. The color scale indicates the voltage measured in the sensor’s receive coil. SAIC inverted this data using a geophysical model. Figure 23(b) shows the data that the final estimated target would produce, based on the model. The estimated data in Figure 23(b) closely matches the collected data in Figure 23(a), giving confidence in the target’s final estimated characteristics [36].

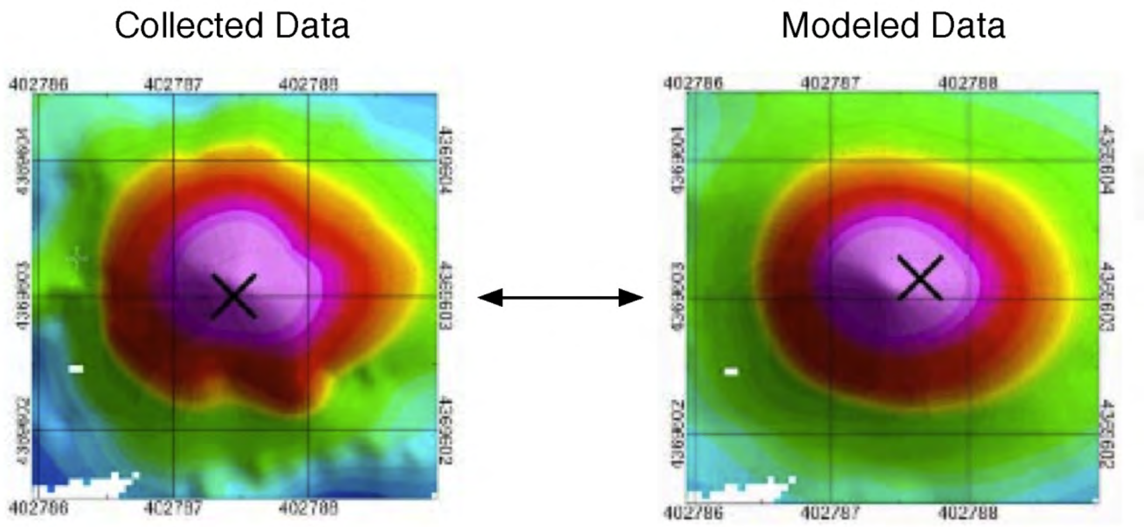


Figure 23: Geophysical inversion. Left: The EM61-Mk2 data collected for one anomaly at the former Camp Butner. The color scale indicates the voltage measured in the sensor’s receive coil, related to the strength of the secondary electromagnetic field induced in the buried target. The data were inverted to estimate the characteristics of the target. Right: The data that the final modeled target would produce, based on a geophysical model. The final modeled data closely matches the collected data, giving confidence in the target’s final estimated characteristics. Taken from [36].

Different teams used different software packages for inversion. For some of its analyses, SAIC used custom-built software written in Interactive Data Language (IDL). The IDL software assumed only one dipole source per anomaly [37]. For other analyses, SAIC used the UX-Analyze package of Oasis montaj. This software called a custom-built module written in IDL that used an iterative process to estimate the number of dipole sources per anomaly [37]. (NAEVA and Parsons also used UX-Analyze for their retrospective analyses. While Parsons configured the software to assume multiple sources per anomaly, the NAEVA configurations assumed both a single source and multiple sources per anomaly [44][46][47].) CH2M HILL and Sky used UXOLab for inversion, a proprietary software package written by the University of British Columbia. UXOLab can also assume one or multiple dipole sources per anomaly [42][49]. Geometrics also used its own software; this software assumed a single dipole source [52][55]. SIG used custom-built software to perform inversions for its retrospective analyses [35]. Each of these software packages is based on a similar geophysical model, as they all assume dipole sources. They differ, though, in their routines for calculating match scores, finding an optimum solution, and adjusting the characteristics of the modeled target.

The final team, Dartmouth, used custom-built software employing much more advanced models that do *not* assume dipole sources [28][56][57][58][59]. These models were inspired by targets with nonhomogeneous material compositions situated in the near field of the sensor. The models are also appropriate for analyzing overlapping anomalies generated by multiple, closely spaced anomalies. Although overlapping anomalies can constitute only a fraction of the anomalies seen on a site, they are often the most challenging anomalies to classify. For example, at the former Camp San Luis Obispo, frequent classification errors resulted from the effects of multiple targets situated within the sensor field of view [34].

Despite their differences, all models can be used to estimate the intrinsic and extrinsic characteristics of the buried targets. Extrinsic characteristics include a target’s location (relative easting and northing coordinates with respect to the sensor platform), as well as its depth and orientation. These characteristics can be used to help guide the recovery of buried targets, as was done when creating the recovery list during the collection of ground truth. In contrast, intrinsic characteristics are related to the physical properties of the target (e.g., size, shape, material composition, wall thickness, etc.) regardless of where or how the target is situated. Successful classification exploits the known differences in the intrinsic characteristics between TOIs and Non-TOIs.

Geophysical models can be used to estimate many intrinsic characteristics of a buried target. For example, the models can be used to estimate the polarizability of the buried target along each of its three major axes at each time gate, often denoted $\beta_1(t_i)$, $\beta_2(t_i)$, and $\beta_3(t_i)$ for a particular time gate t_i. Figure 24(a) and (b) show photographs and plots of the polarizability curves for an M48 fuze (TOI) and a piece of scrap metal (Non-

TOI), as estimated from the MetalMapper data. At a given time gate t_i, the amplitudes of the polarizabilities provide a measure of the target's size, while ratios of the polarizabilities provide a measure of the target's aspect ratio or shape. TOIs tend to be bodies of revolution with one large axis and two equal, smaller axes. Thus, TOIs usually exhibit one large polarizability, along with two equal, smaller ones. In contrast, Non-TOIs such as cultural debris may be small and are not often bodies of revolution. Therefore, Non-TOIs often exhibit three different polarizabilities. Attempts can also be made to estimate the decays of the polarizabilities over time (denoted τ_1, τ_2, and τ_3). TOIs tend to be ferrous in composition and have thicker walls than Non-TOIs, leading to slower decay rates. Decay rates can be difficult to estimate using dynamic data, however, because the time gates of the dynamic sensors do not extend far enough in time to capture the late differences in decay rates between TOIs and Non-TOIs. More advanced instruments, such as the TEMTADS and the MetalMapper in static mode, sample the received signal at later time gates than the dynamic instruments. Models applied to the static data can therefore produce more accurate estimates of τ_1, τ_2, and τ_3.

2. Data-Driven Characterizations

NAEVA was the only team that did not use a geophysical model to characterize the buried targets [5][18][25][26][47][48]. Instead, this team measured the characteristics of the EM61-Mk2 data using functions embedded in the UX-Detect and UX-Process modules of the Oasis montaj software. An example of these measurements includes the peak amplitude of the received signal, which is a very rough estimate of the size of the buried target. Target depth can confound this estimate, however, because a large, deep target can give a similar amplitude as a small, shallow target. In addition, NAEVA also estimated the decay rate of the signal amplitude, similar to the polarizability decay rates (τ_1, τ_2, and τ_3) estimated by the other teams using geophysical models.

3. Selecting Features for Classification

Each team decided which target characteristics to use for classification. Some teams, such as Sky, input the full polarizability curves, $\beta_1(t_i)$, $\beta_2(t_i)$, and $\beta_3(t_i)$ for all time gates t_i, into their classification algorithms [50][51]. Other teams measured features from the curves (or from the data themselves) that they believed were most likely to exploit the known differences between TOIs and non-TOIs. Only these measured features were input into the classification algorithms. For example, in one of its analyses, NAEVA chose a single feature, the decay rate of the peak amplitude of the detected anomaly (related to the material composition and wall thickness of the buried target) [25]. This resulted in a one-dimensional classification analysis. Sky performed two-dimensional analyses, based on the sum of the polarizabilities at the first time gate $\Sigma\beta_i(t_1)$ (related to the target's size), as well as the ratio of the primary polarizability at the first to the fourth time gates $\beta_1(t_1)/\beta_1(t_4)$ (one way to calculate τ_1, related to the target's wall thickness) [22][50][51].

Figure 25 shows a plot of these two features with respect to each other. Points shaded in red, blue, and yellow were estimated from anomalies with TOI ground-truth labels; points shaded in black were estimated from Non-TOI anomalies. The TOI and Non-TOI anomalies occupy somewhat different regions of feature space, suggesting that these two features could be used to classify the anomalies.

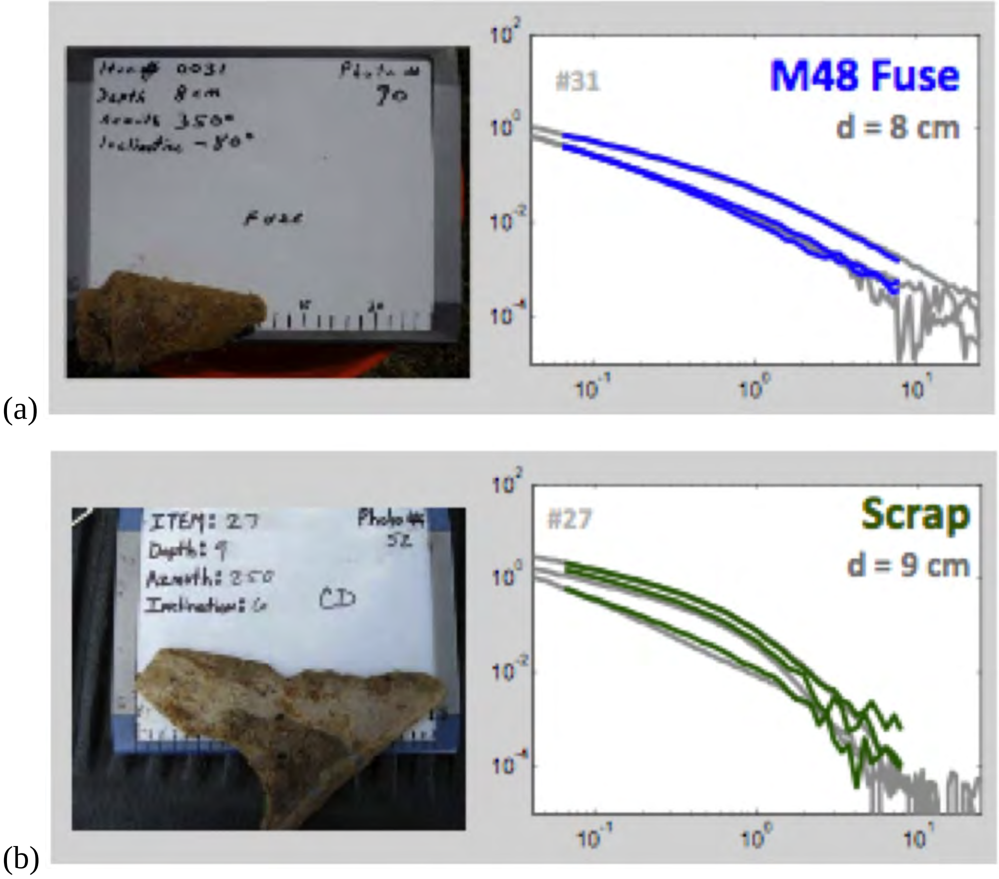


Figure 24: Photographs and estimated polarizability curves for (a) an M48 fuze (TOI) and (b) a piece of scrap metal (Non-TOI), as inverted from MetalMapper static data. The TOI has one large polarizability along with two smaller and relatively equal polarizabilities, indicating a body of revolution with one large axis and two equal, smaller axes. The TOI's polarizabilities decay at a slow rate, indicating a ferrous composition and thick walls. In contrast, the Non-TOI has three polarizabilities of different amplitudes, all of which decay more quickly. This indicates a nonferrous, thin-walled object that is not a body of revolution. Taken from [51].

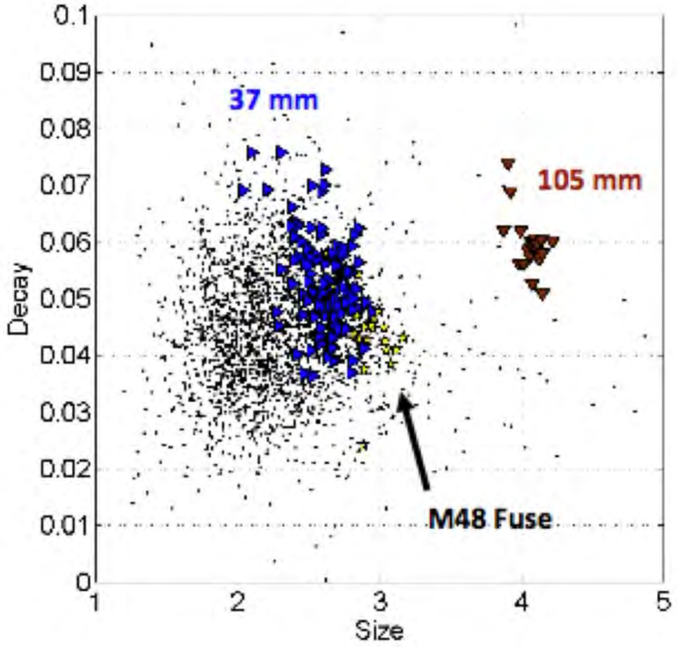


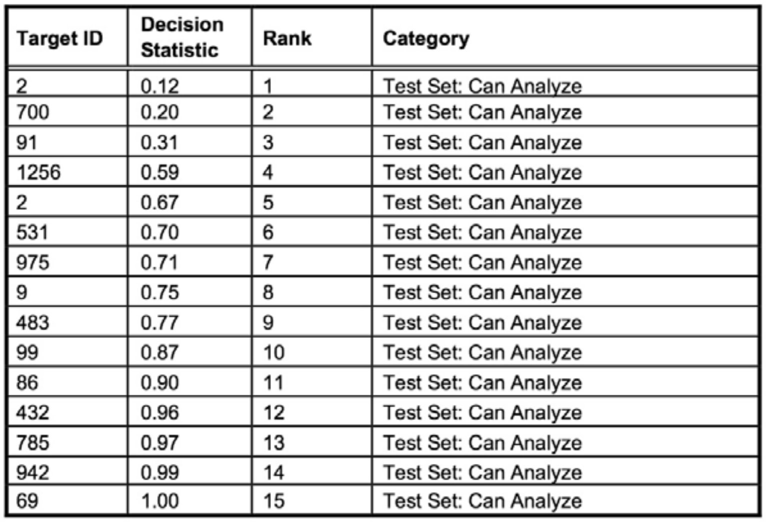
Figure 25: A two-dimensional plot of features estimated from the MetalMapper dynamic data collected for all anomalies in the demonstration area. The horizontal axis indicates the sum of the polarizabilities of the buried target (related to size), while the vertical axis indicates the decay rate of the principal polarizability (related to material composition and wall thickness). In general, the red, blue, and yellow TOI anomalies cluster in different regions of feature space than the black Non-TOI anomalies, indicating that these two features could be used for classification. Taken from [51].

E. Classifying the Buried Targets

The classification analyst teams designed algorithms to classify the anomalies based on the selected features. Different teams used different algorithms. NAEVA used rule-based decision trees implemented in custom-built software [5][18][25][26][47][48]. SAIC used a library-matching algorithm built in to UX-Analyze to compare each anomaly to a library of known TOIs and Non-TOIs [2][10][11][12][13][14][38][39]. (NAEVA and Parsons also used the UX-Analyze library-matching algorithm in their retrospective analyses [6][15][16][44][45][46][47].) Sky used a library-matching algorithm for some of its analyses and a statistical classifier for other analyses. Both algorithms were implemented in UXOLab [22][50][51]. A final Sky analysis was based on the library-matching algorithm coupled with input from a human expert. CH2M HILL did the same [42]. Dartmouth and SIG used statistical classifiers implemented in custom-built software [35][57][58]. Geometrics performed three analyses, all implemented in custom-built software, as well [55]. One analysis was based on a neural network (a type of statistical classifier), a second was based on a neural network coupled with a set of rules, and a third was based on a neural network, rules, *and* a library-matching algorithm [52].

Despite their differences, most of the classification algorithms functioned in similar ways: First, the features estimated for each anomaly were input into the algorithm. The algorithm output the anomaly's likelihood of being a Non-TOI or another similar decision statistic. The classification analyst team used the decision statistic to create a ranked anomaly list, the final product of the classification analysis. A ranked anomaly list is an ordered list of all anomalies in the test set. Ranked anomaly lists were constructed in three distinct steps.

In the first step, the analysts ranked the test set anomalies assigned to the "Can Analyze" group. The ranks were based on the decision statistics estimated by the classification algorithm. The anomalies were ordered on the list from most to least likely to be Non-TOIs, as shown in Figure 26. That is to say, the first anomaly on the list (with a rank of 1) was the anomaly deemed most likely to be a Non-TOI, and the last anomaly on the list was the one deemed most likely to be a TOI.



Target ID	Decision Statistic	Rank	Category
2	0.12	1	Test Set: Can Analyze
700	0.20	2	Test Set: Can Analyze
91	0.31	3	Test Set: Can Analyze
1256	0.59	4	Test Set: Can Analyze
2	0.67	5	Test Set: Can Analyze
531	0.70	6	Test Set: Can Analyze
975	0.71	7	Test Set: Can Analyze
9	0.75	8	Test Set: Can Analyze
483	0.77	9	Test Set: Can Analyze
99	0.87	10	Test Set: Can Analyze
86	0.90	11	Test Set: Can Analyze
432	0.96	12	Test Set: Can Analyze
785	0.97	13	Test Set: Can Analyze
942	0.99	14	Test Set: Can Analyze
69	1.00	15	Test Set: Can Analyze

Figure 26: A sketch of a ranked anomaly list in its first stage of construction. "Can Analyze" anomalies in the test set are ranked according to their estimated likelihoods of being Non-TOIs or a similar decision statistic. The first anomaly on the list (with a rank of 1) was the anomaly deemed to be most likely to be a Non-TOI. The last anomaly was the one deemed most likely to be a TOI.

In the second step, the analysts further classified the anomalies on the list. As shown in Figure 27, the "Can Analyze" anomalies were further separated into three classes, "Likely TOI," "Cannot Decide," "Likely Non-TOI." Analysts chose the boundaries between the different classes by assessing the information available in the training set. The boundary between the green "Likely Non-TOI" and yellow "Cannot Decide" classes was the most important boundary because it constituted the analyst's "don't dig threshold" (see Figure 28).

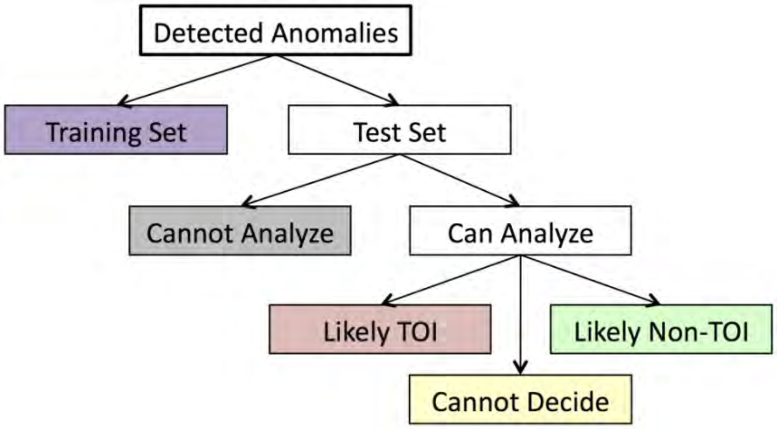
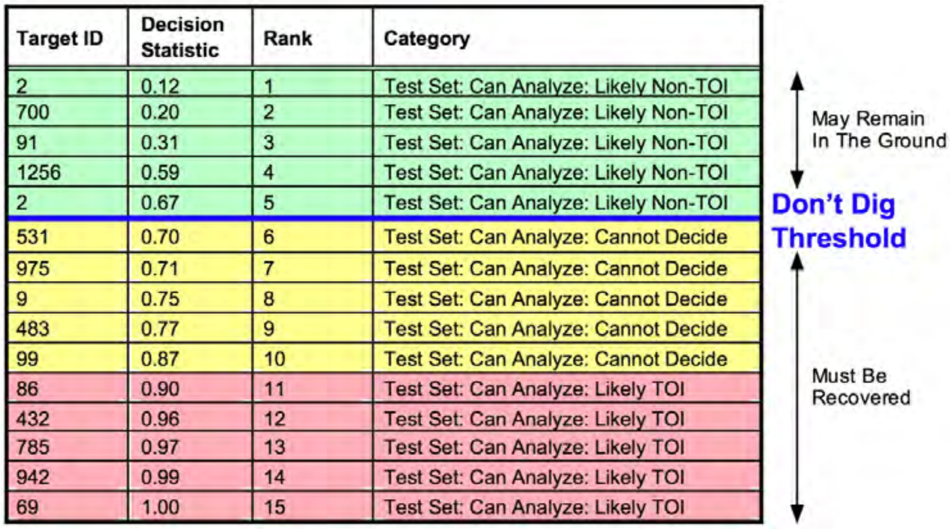


Figure 27: Each classification analyst team further separated the "Can Analyze" test set anomalies into "Likely TOI," "Cannot Decide," and "Likely Non-TOI" categories.

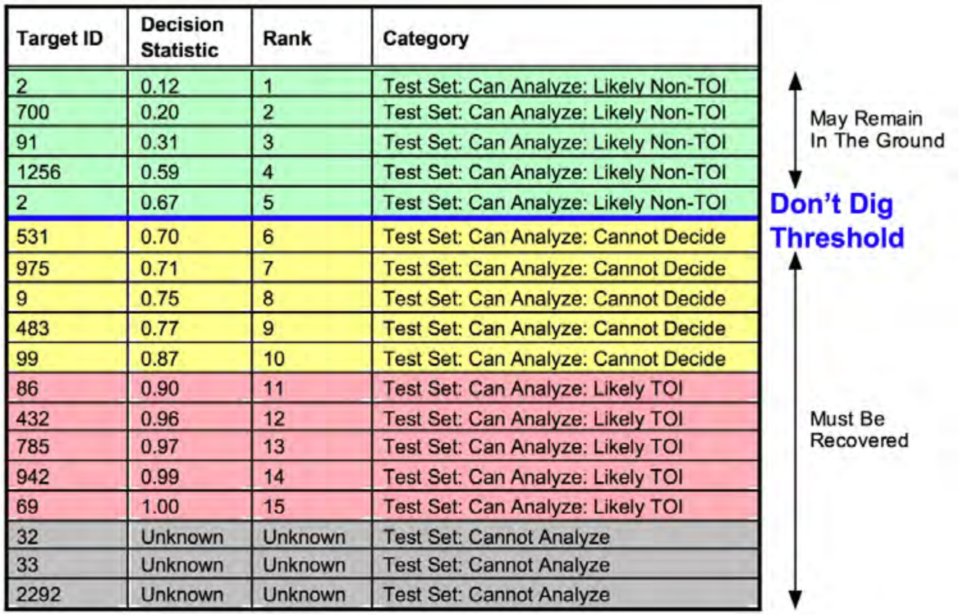


Target ID	Decision Statistic	Rank	Category
2	0.12	1	Test Set: Can Analyze: Likely Non-TOI
700	0.20	2	Test Set: Can Analyze: Likely Non-TOI
91	0.31	3	Test Set: Can Analyze: Likely Non-TOI
1256	0.59	4	Test Set: Can Analyze: Likely Non-TOI
2	0.67	5	Test Set: Can Analyze: Likely Non-TOI
531	0.70	6	Test Set: Can Analyze: Cannot Decide
975	0.71	7	Test Set: Can Analyze: Cannot Decide
9	0.75	8	Test Set: Can Analyze: Cannot Decide
483	0.77	9	Test Set: Can Analyze: Cannot Decide
99	0.87	10	Test Set: Can Analyze: Cannot Decide
86	0.90	11	Test Set: Can Analyze: Likely TOI
432	0.96	12	Test Set: Can Analyze: Likely TOI
785	0.97	13	Test Set: Can Analyze: Likely TOI
942	0.99	14	Test Set: Can Analyze: Likely TOI
69	1.00	15	Test Set: Can Analyze: Likely TOI

Figure 28: A sketch of a ranked anomaly list in its second stage of construction. "Can Analyze" anomalies in the test set are further separated into three categories based on their estimated likelihoods of being Non-TOIs, or another similar decision statistic. Anomalies classified as "Likely Non-TOI," "Cannot Decide," and "Likely TOI" are colored in green, yellow, and red, respectively. A thick blue line indicates the don't dig threshold, the boundary between the green and yellow categories. In a UXO remediation project, the recovery team would begin recovering targets from the bottom of the list and work its way up until it reached the don't dig threshold. Targets above threshold could be dug using more relaxed safety-oriented precautions or could simply remain in the ground.

The don't dig threshold informs stakeholders which anomalies must be excavated. In a UXO remediation project, the recovery team could be instructed to recover the most dangerous targets first, those that produced anomalies classified as "Likely TOI." The recovery team would have to err on the side of caution, though, and also recover targets that produced the anomalies classified as "Cannot Decide." At that point, stakeholders could instruct the recovery team to leave presumably innocuous targets in the ground, those that produced anomalies classified as “Likely Non-TOI.” Stakeholders could instruct the team to cease recovery upon reaching the don’t dig threshold, as evidence exists that the buried targets consist of clutter only. Alternatively, the team could be instructed to continue digging but with more relaxed safety-oriented precautions.

In the third and final step of creating a ranked anomaly list, the analysts focused on the test set anomalies in the “Cannot Analyze” group. In a UXO remediation project, the recovery team must err on the side of caution and recover all possibly dangerous targets, including those that produced anomalies that could not be analyzed. This means that all “Cannot Analyze” anomalies must be inserted into the ranked anomaly list at a point below the don’t dig threshold. ESTCP instructed the analysts to append the “Cannot Analyze” anomalies to the bottom of the ranked anomaly list, as shown in Figure 29. The “Cannot Analyze” anomalies were arranged in no particular order with respect to each other because, by definition, no further information could be learned about them, including their likelihoods of being Non-TOIs.



Target ID	Decision Statistic	Rank	Category
2	0.12	1	Test Set: Can Analyze: Likely Non-TOI
700	0.20	2	Test Set: Can Analyze: Likely Non-TOI
91	0.31	3	Test Set: Can Analyze: Likely Non-TOI
1256	0.59	4	Test Set: Can Analyze: Likely Non-TOI
2	0.67	5	Test Set: Can Analyze: Likely Non-TOI
531	0.70	6	Test Set: Can Analyze: Cannot Decide
975	0.71	7	Test Set: Can Analyze: Cannot Decide
9	0.75	8	Test Set: Can Analyze: Cannot Decide
483	0.77	9	Test Set: Can Analyze: Cannot Decide
99	0.87	10	Test Set: Can Analyze: Cannot Decide
86	0.90	11	Test Set: Can Analyze: Likely TOI
432	0.96	12	Test Set: Can Analyze: Likely TOI
785	0.97	13	Test Set: Can Analyze: Likely TOI
942	0.99	14	Test Set: Can Analyze: Likely TOI
69	1.00	15	Test Set: Can Analyze: Likely TOI
32	Unknown	Unknown	Test Set: Cannot Analyze
33	Unknown	Unknown	Test Set: Cannot Analyze
2292	Unknown	Unknown	Test Set: Cannot Analyze

Figure 29: A sketch of a ranked anomaly list in its third and final stage of construction. “Cannot Analyze” anomalies in the test set have been appended to the bottom of the list in no particular order with respect to each other.

At first glance, the order of the anomalies on the ranked anomaly list may seem reversed. In real remediation projects, anomaly lists are often ordered in the reverse, with anomalies at the top of the list most likely to be TOIs and those at the bottom most likely to be Non-TOIs. In this demonstration, though, the reverse is true. ESTCP intentionally chose to order the ranked anomaly lists this way to remind the UXO community that the purpose of UXO classification is to correctly identify *Non-TOIs*, such that only those targets would be left in the ground (or recovered using more relaxed safety-oriented precautions). To that end, the threshold separating the green "Likely Non-TOI" anomalies from all other anomalies is referred to as the "*don't* dig threshold." All anomalies surpassing this threshold do *not* have to be dug.

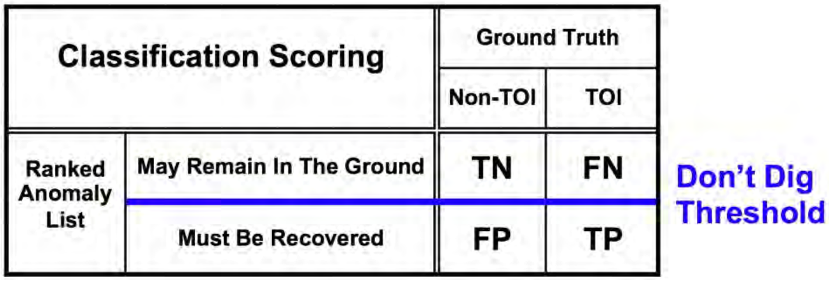
Each classification analyst team created one ranked anomaly list for each one of its analyses. A total of 54 ranked anomaly lists from 8 different teams was submitted to ESTCP and scored by IDA. (Twenty-two of the 54 lists were the products of retrospective analyses, submitted after the demonstration was complete and ground truth had been released to the public.)

6. Scoring the Ranked Anomaly Lists

One of the main goals of this demonstration was to evaluate the performance of different types of UXO classification analyses. The final products of the analyses were ranked anomaly lists. IDA scored each ranked anomaly list by comparing it to ground truth, calculating performance metrics based on this comparison, and plotting the metrics with respect to each other to form a classification performance curve, similar to a receiver-operating characteristic (ROC) curve. This section describes how these curves were formed.

Performance metrics were calculated for each ranked anomaly list. Counts were made of the number of true-positive (TP), false-positive (FP), true-negative (TN), and false-negative (FN) anomalies on the list. As shown in Figure 30, these counts were based on (1) the anomalies' ground-truth labels and (2) where the anomalies fell on the ranked anomaly list with respect to the don't dig threshold:

- Number of TOIs dug = TP
- Number of Non-TOIs dug = FP
- Number of TOIs not dug = FN
- Number of Non-TOIs not dug = TN



Classification Scoring		Ground Truth	
		Non-TOI	TOI
Ranked Anomaly List	May Remain In The Ground	TN	FN
	Must Be Recovered	FP	TP

Don't Dig Threshold

Figure 30: Metrics used to score the classification performance of a ranked anomaly list. True-positive (TP) and false-positive (FP) anomalies are those that correctly and incorrectly fell below the don't dig threshold, respectively—their buried targets must be dug. These metrics were plotted with respect to each other to form a classification performance curve. True-negative (TN) and false-negative (FN) anomalies are those that correctly and incorrectly rose above the don't dig threshold, respectively—their buried targets could remain in the ground.

IDA included each anomaly on the ranked anomaly list into one of these four counts. There were some nuances to this process, however. When recovering all buried targets to produce ground truth, Parsons Inc. reported that two anomalies, #2409 and

#3721, had been caused by the same piece of munitions debris and must therefore “share” the same buried target. To avoid double-counting this target during scoring, IDA included in the tallies only that anomaly classified as most likely to be TOI, ignoring the other anomaly. An identical situation occurred with two other anomalies caused by the same piece of munitions debris, #1019 and #2144.

The four counts of TP, FP, FN, and TN behave in the following ways:

- Total number of TOIs (dug and not dug) = TP + FN
- Total number of Non-TOIs (dug and not dug) = FP + TN
- Total number of targets dug (TOIs and Non-TOIs) = TP + FP
- Total number of targets not dug (TOIs and Non-TOIs) = FN + TN

In many general classification problems, summary metrics are often calculated based on the four counts. For example, the probability of detection (Pd) and the probability of false alarm (Pfa) are often calculated as Pd = TP / (TP + FN) and Pfa = FP / (FP + TN). Note, however, that the Pd calculated for classification problems is different than the Pd calculated for detection problems. For example, Pd was calculated during detection scoring to estimate the percentage of TOIs that were correctly detected. Here, though, Pd describes a slightly different metric related to classification performance: the percentage of all detected TOIs that were correctly dug. Similarly, Pfa describes the percentage of all detected *Non*-TOIs that were *incorrectly* dug. ESTCP considered using these metrics in this demonstration. After further consideration, though, the decision was made to use different summary metrics instead.

This demonstration used the raw counts of TP and FP in place of the summary metrics Pd and Pfa. Stakeholders in UXO remediation projects are often interested in the total number of targets dug (TP + FP) because this number drives much of the cost of the project. TP represents the number of digs to recover TOIs. This count can be easily multiplied by the cost per dig to arrive at the necessary dig costs of UXO remediation. (These dig costs are necessary because TOIs must be recovered.) In contrast, FP represents the number of digs to recover Non-TOIs. This value can be easily translated into the *unnecessary* dig costs of UXO remediation. (These digs costs are unnecessary, since the Non-TOIs could have remained safely in the ground.) Stakeholders would like as many TOIs to be recovered as possible. Therefore, the number of TOI digs (TP) should be as near as possible to the total number of TOIs (FN + TP). (In other words, FN should be as near as possible to zero.) Stakeholders also want to reduce unnecessary costs as much as possible. Therefore, the number of Non-TOI digs (FP) should be as low as possible.

IDA plotted the number of TOI digs (TP) versus the number of Non-TOI digs (FP) for each ranked anomaly list. Figure 31 shows a sketch of such a plot. The vertical axis ranges from zero to the maximum possible number of TOI digs, that is, the total number of TOIs (TP + FN). The horizontal axis ranges from zero to the maximum possible number of Non-TOI digs, that is, the total number of Non-TOIs (FP + TN). The plotted point (large blue dot) illustrates the classification performance that results when the analyst's don't dig threshold is applied to the ranked anomaly list. A vertical gray bar passing through the large blue dot estimates the 95% confidence interval through the point.

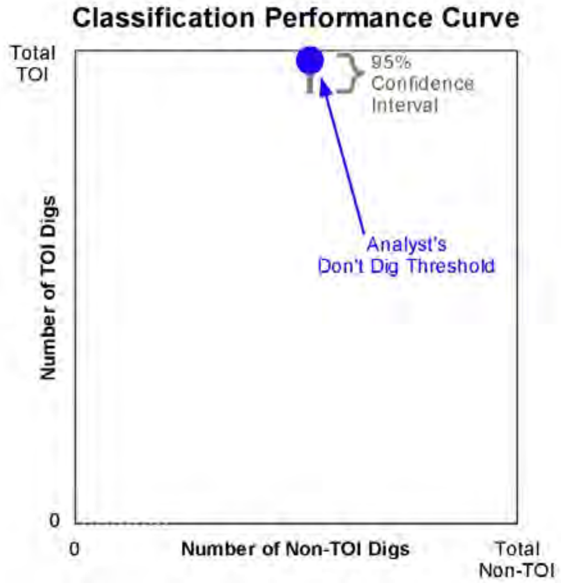
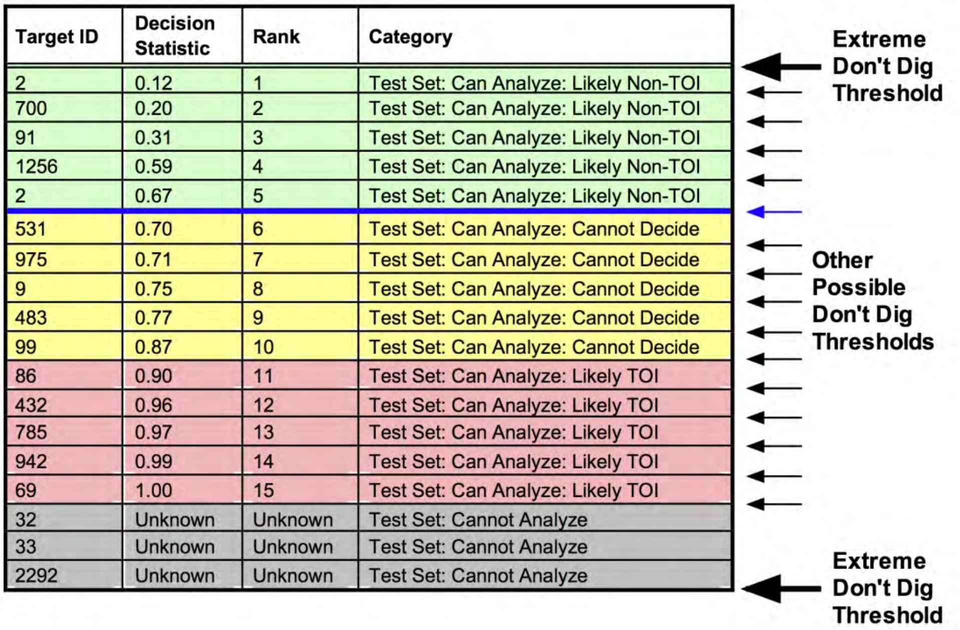


Figure 31: A sketch illustrating the results of applying the analyst's don't dig threshold to a ranked anomaly list. The large blue dot indicates the number of TOI versus Non-TOI digs (TP versus FP) resulting from the analyst's don't dig threshold. The vertical and horizontal axes range from zero to the maximum possible number of TOIs and Non-TOIs, respectively. The vertical gray bar passing through the large blue dot estimates the 95% confidence interval around the number of TOI digs (TP).

This plot was used to revisit the choice of don't dig threshold. The analysts had prospectively chosen one particular don't dig threshold to apply to the ranked anomaly list. Other don't dig thresholds could have been chosen instead. This was illustrated by retrospectively applying all possible don't dig thresholds to the ranked anomaly list, one by one, as shown in Figure 32. In the most extreme cases, the don't dig threshold could have been placed at the top or bottom of the list. In between the two extremes, the don't dig threshold could have been placed in the green, yellow, or red parts of the list.

For each possible don't dig threshold, TP, FP, FN, and TN were re-counted and TP (the number of TOI digs) was re-plotted against FP (the number of Non-TOI digs). Figure 33 shows a sketch of this plot. Each point (black dot) corresponds to one possible don't dig threshold. Together, the points form a classification performance curve. This curve is similar to the ROC curves of Pd versus Pfa that are often used in general classification problems. Vertical gray bars indicate the 95% confidence intervals through the points.



Target ID	Decision Statistic	Rank	Category
2	0.12	1	Test Set: Can Analyze: Likely Non-TOI
700	0.20	2	Test Set: Can Analyze: Likely Non-TOI
91	0.31	3	Test Set: Can Analyze: Likely Non-TOI
1256	0.59	4	Test Set: Can Analyze: Likely Non-TOI
2	0.67	5	Test Set: Can Analyze: Likely Non-TOI
531	0.70	6	Test Set: Can Analyze: Cannot Decide
975	0.71	7	Test Set: Can Analyze: Cannot Decide
9	0.75	8	Test Set: Can Analyze: Cannot Decide
483	0.77	9	Test Set: Can Analyze: Cannot Decide
99	0.87	10	Test Set: Can Analyze: Cannot Decide
86	0.90	11	Test Set: Can Analyze: Likely TOI
432	0.96	12	Test Set: Can Analyze: Likely TOI
785	0.97	13	Test Set: Can Analyze: Likely TOI
942	0.99	14	Test Set: Can Analyze: Likely TOI
69	1.00	15	Test Set: Can Analyze: Likely TOI
32	Unknown	Unknown	Test Set: Cannot Analyze
33	Unknown	Unknown	Test Set: Cannot Analyze
2292	Unknown	Unknown	Test Set: Cannot Analyze

Extreme Don't Dig Threshold
Other Possible Don't Dig Thresholds
Extreme Don't Dig Threshold

Figure 32: A sketch of a ranked anomaly list, indicating all possible don't dig thresholds. All anomalies would have fallen below the most extreme don't dig threshold at the very top of the list. In contrast, no anomalies would have fallen below the other extreme threshold at the very bottom of the list. In between the two extremes, there is one possible don't dig threshold for each unique rank in the green, yellow, and red parts of the list. The don't dig threshold could not have been placed in the gray part of the list because "Cannot Analyze" anomalies share the same rank (which happens to be "Unknown").

A two-step process was used to estimate the 95% confidence interval through every point. First, for each don't dig threshold, the exact binomial distribution was used to estimate CI_{Pd}, the 95% confidence interval around Pd, where Pd = TP / (TP + FN). In the second step, CI_{Pd} was converted into CI_{TP}, the 95% confidence interval around TP. Since TP is the numerator of Pd, CI_{TP} was calculated by multiplying CI_{Pd} by (TP + FN), the denominator of Pd.

The 95% confidence interval was calculated independently for each possible don't dig threshold, without any adjustments for multiple comparisons. This means that one *cannot* infer that 95 times out of 100, every point on the curve will simultaneously lie within its own 95% confidence interval. That is, one can*not* infer that 95 times out of 100, the entire curve will lie within the band generated by "smearing" the individual confidence intervals [43]. However, the confidence intervals are still a useful means of assessing the confidence around an individual point, such as when comparing the same point on two different curves. For example, if the vertical gray bar passing through the dark blue dot on one curve (representing one analyst's don't dig threshold) does not overlap vertically with the bar passing through the dark blue dot on another curve (representing another analyst's don't dig threshold), then evidence exists that the analysts' thresholds are statistically different from each other. On the other hand, if the two confidence intervals do indeed overlap, then there is no evidence that the two thresholds are statistically different [43].

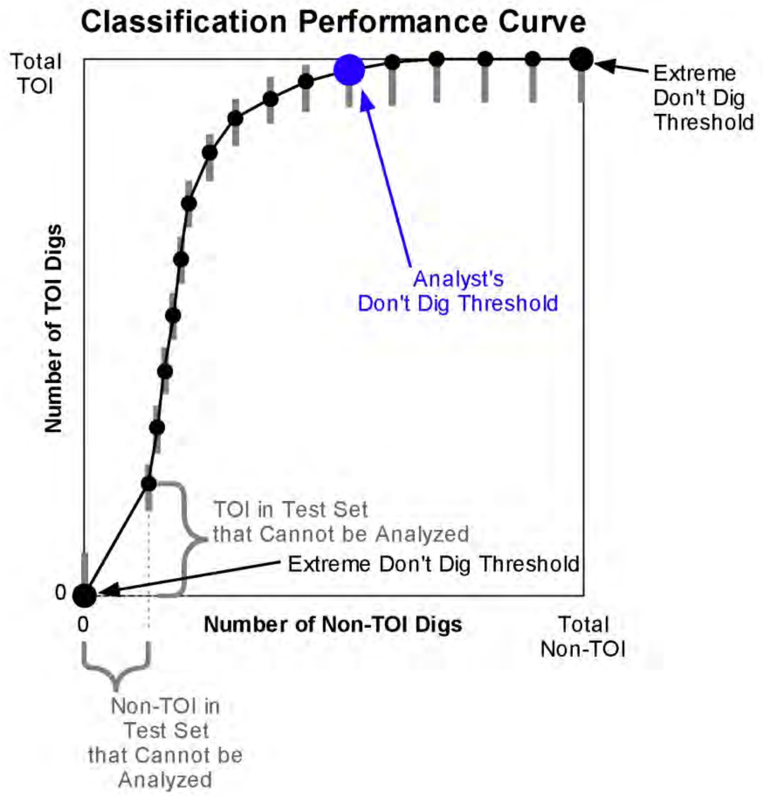


Figure 33: A sketch of a classification performance curve, illustrating the results of applying all possible don't dig thresholds to a ranked anomaly list. Small black dots indicate the number of TOI digs (TP) versus the number of Non-TOI digs (FP) from the different don't dig thresholds. The black dot in the upper right corner represents the extreme case in which the don't dig threshold is placed at the very top of the list, such that all anomalies fall below. The black dot in the lower left corner represents the other extreme case in which the don't dig threshold is placed at the bottom of the list, such that no anomalies fall below. The gap between the lower left corner and the next closest point represents the "Cannot Analyze" anomalies.

As with ROC curves, the points on a classification performance curve are not always equally spaced. All points on the curve lying between the lower left and upper right corners represent possible don't dig thresholds, with one don't dig threshold per each *unique* rank on the ranked anomaly list. Not all ranks are unique, however. Some anomalies on the list may share the same rank if they were considered equally likely to be Non-TOIs. Therefore, these anomalies must fall together either above or below any given don’t dig threshold; a don’t dig threshold cannot be placed between them. These groups of identically ranked anomalies lead to gaps between points in the curve.

The gap is particularly noticeable for the anomalies deemed “Cannot Analyze” and appended to the bottom of the ranked anomaly list. By definition, the data collected for these anomalies could not be analyzed. Specifically, the buried target could not be characterized, and the classification algorithm could not estimate the anomalies’ likelihoods of being Non-TOIs nor any other decision statistic. Therefore, all “Cannot Analyze” anomalies share the rank of “Unknown.” Because they share the same rank, a don’t dig threshold cannot be placed between them. This causes a gap between the lower left corner of the plot and the next closest point, as shown in Figure 33. This next point represents the case where the don’t dig threshold is placed directly between the “Can Analyze” and “Cannot Analyze” anomalies. Only the “Cannot Analyze” anomalies fall below this don’t dig threshold. In this case, if Y of the TOI anomalies are deemed “Cannot Analyze” and therefore fall below the don’t dig threshold, then the number of TOI digs (TP) is equal to Y. Similarly, if X of the Non-TOI anomalies are deemed “Cannot Analyze” and therefore fall below threshold, then the number of Non-TOI digs (FP) is equal to X.

Figure 34 shows a similar sketch of the curve, this time with the points plotted in color. The analysts had separated the “Can Analyze” anomalies into three categories on the ranked anomaly list. A point on the curve was colored in red, yellow, or green if the anomaly directly above the corresponding don’t dig threshold was classified into the “Likely TOI,” “Cannot Decide,” or “Likely Non-TOI” category, respectively. By definition, the analyst’s don’t dig threshold (large blue dot) lies between the green and yellow points, the boundary between the “Cannot Decide” and “Likely Non-TOI” categories on the ranked anomaly list.

Each classification performance curve was then examined to identify the best possible choice of don’t dig threshold. Choosing the don’t dig threshold is a critical step in UXO classification. A don’t dig threshold placed near the top of the ranked anomaly list will lead to a large number of TOI digs (a desirable outcome) but also a large number of Non-TOI digs (an undesirable outcome). Conversely, placing the don’t dig threshold near the bottom of the list will lead to a small number of TOI digs (an undesirable outcome) but also a small number of Non-TOI digs (a desirable outcome). The best don’t dig threshold lies somewhere in between.

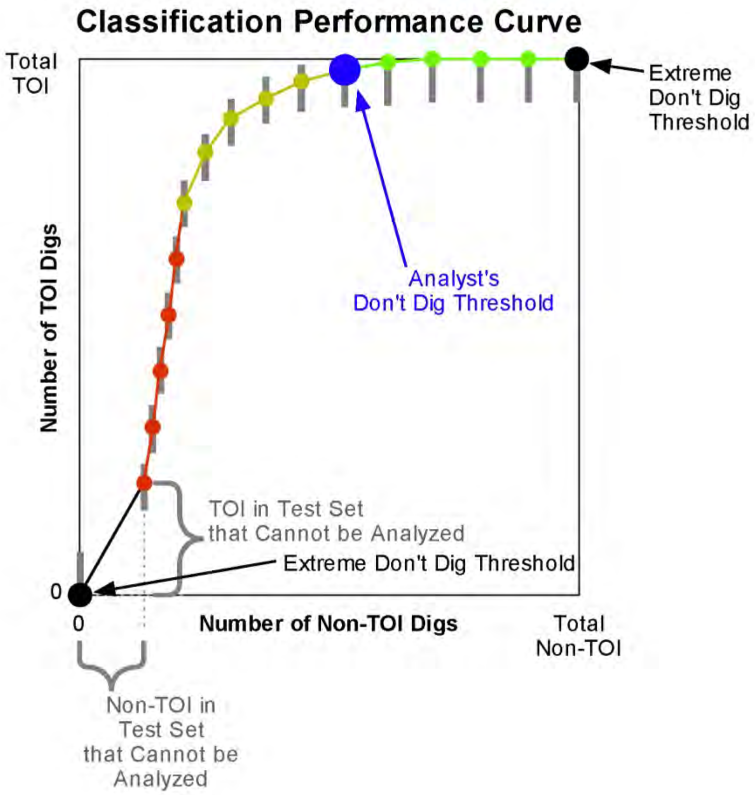


Figure 34: A sketch of a classification performance curve, noting into which categories the possible don't dig thresholds fell on the ranked anomaly list. Points colored in red, yellow, and green correspond to don't dig thresholds that fell in the "Likely TOI," "Cannot Decide," and "Likely Non-TOI" categories, respectively. By definition, the analyst's don't dig threshold lies between the yellow and green points.

IDA retrospectively identified the "best" don't dig threshold for each ranked anomaly list. The "best" threshold was defined as that which would have resulted in the smallest number of Non-TOI digs (i.e., minimum FP) while the number of TOI digs was held at its maximum possible value (i.e., Pd = 100%). This don't dig threshold would have minimized the cost of recovering targets while leaving no TOIs in the ground. All TOI anomalies would have been correctly classified, with the fewest Non-TOI anomalies incorrectly classified. The large light blue dot in Figure 35 indicates this "best" don't dig threshold.

A second don't dig threshold was also identified retrospectively. This threshold would have resulted in the smallest number of Non-TOI digs (i.e., minimum FP) while the number of TOI digs was held at no less than 95% of its maximum possible value (i.e., Pd ≥ 95%). At least 95% of the TOI anomalies would have been correctly classified, with the fewest Non-TOI anomalies incorrectly classified. This don't dig threshold would have minimized the cost of recovering targets while leaving in the ground only 5% of the TOIs (the most difficult to identify). Figure 35 indicates this second don't dig threshold with a large pink dot.

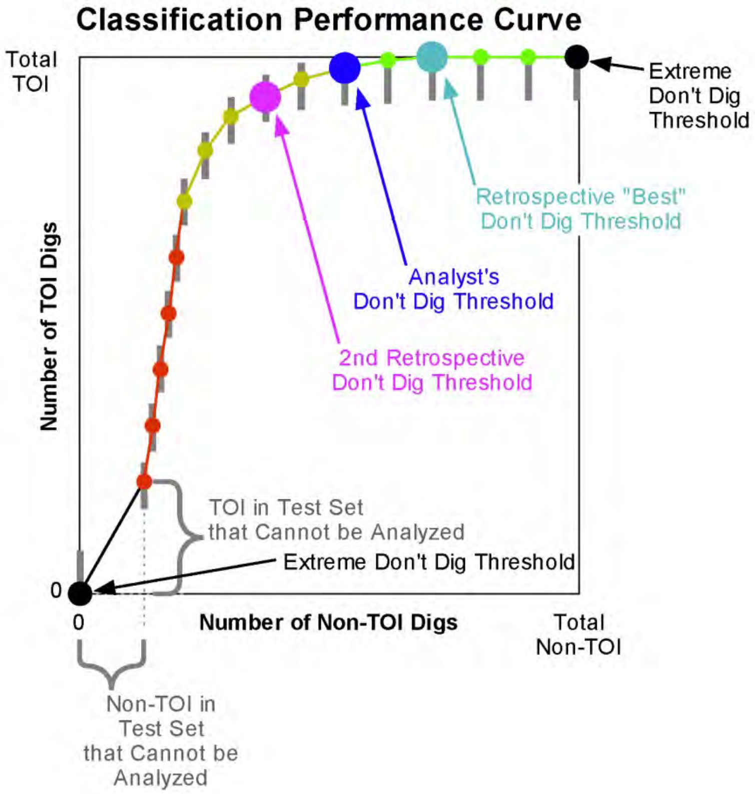
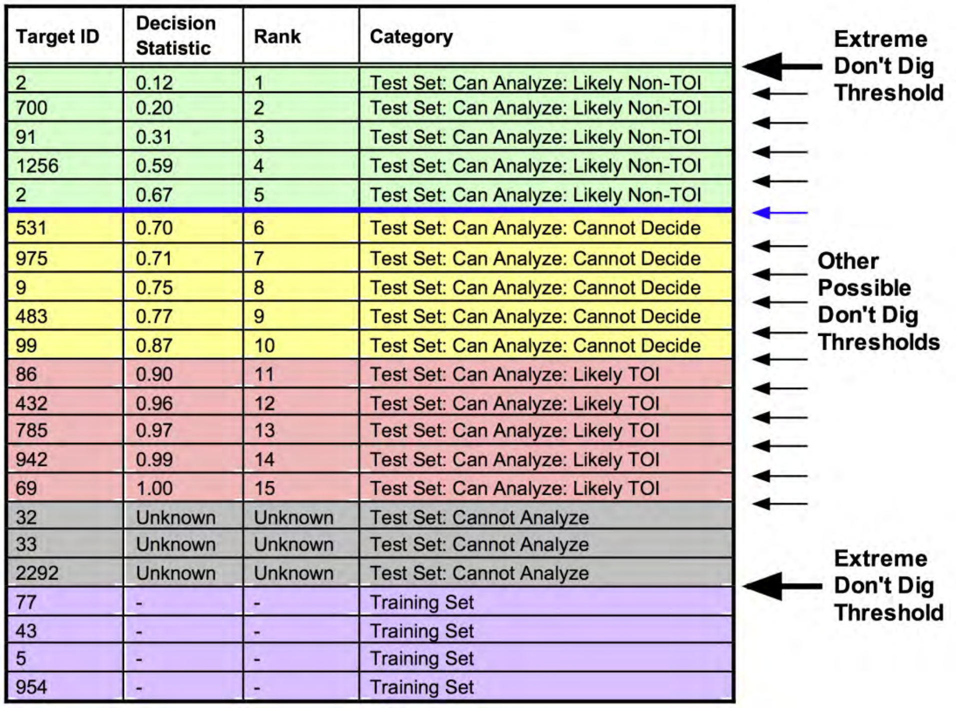


Figure 35: A sketch of a classification performance curve, indicating the retrospectively chosen don't dig thresholds. The large light blue dot represents the "best" don't dig threshold, that which would have minimized the number of Non-TOI digs while the number of TOI digs was held at its maximum value, leaving no TOIs in the ground. The large pink dot represents a second retrospectively chosen threshold, that which would have minimized the number of Non-TOI digs while ensuring that the number of TOI digs was no less than 95% of its maximum possible value, leaving no more than 5% of the TOIs in the ground.

The classification performance curves were then further adjusted to take the training sets into consideration. The purpose of a classification performance curve is to illustrate the performance of a classification analysis, that which includes methods for characterizing the buried targets, classifying the targets based on their characteristics, and optimizing the classification algorithm over a training set. The classification performance curve illustrated in Figure 35 does not take the training set into account, however. This curve is based only on the ranked anomaly list, and the ranked anomaly list consists only of anomalies assigned to the test set. Curves like this can be compared to each other only if they are based on the same test set. In this demonstration, however, different classification analyst teams chose different training and test sets. Test sets differed in both size and character, leading to inherent differences in the TP and FP counts. For example, a classification analysis could have more easily achieved a small number of Non-TOI digs (a low FP) if the test set contained few Non-TOI anomalies in the first place. Similarly, a classification analysis could have exhibited a small number of TOI digs (a low TP) if the test set contained few TOI anomalies in the first place.

To address this issue, the classification performance curves were altered so that they could be compared to each other regardless of which training and test sets were used. In a real remediation project, all targets producing anomalies assigned to the training set must be recovered to obtain ground truth for algorithm optimization. This is true regardless of which don't dig threshold is eventually applied to the ranked anomaly list. Therefore, the training set anomalies were treated as though they had been appended to the very bottom of the ranked anomaly list, below even the most extreme don't dig threshold, as shown in Figure 36. The number of TOI and Non-TOI digs were re-counted and re-plotted to incorporate the training set, as shown in Figure 37.



Target ID	Decision Statistic	Rank	Category
2	0.12	1	Test Set: Can Analyze: Likely Non-TOI
700	0.20	2	Test Set: Can Analyze: Likely Non-TOI
91	0.31	3	Test Set: Can Analyze: Likely Non-TOI
1256	0.59	4	Test Set: Can Analyze: Likely Non-TOI
2	0.67	5	Test Set: Can Analyze: Likely Non-TOI
531	0.70	6	Test Set: Can Analyze: Cannot Decide
975	0.71	7	Test Set: Can Analyze: Cannot Decide
9	0.75	8	Test Set: Can Analyze: Cannot Decide
483	0.77	9	Test Set: Can Analyze: Cannot Decide
99	0.87	10	Test Set: Can Analyze: Cannot Decide
86	0.90	11	Test Set: Can Analyze: Likely TOI
432	0.96	12	Test Set: Can Analyze: Likely TOI
785	0.97	13	Test Set: Can Analyze: Likely TOI
942	0.99	14	Test Set: Can Analyze: Likely TOI
69	1.00	15	Test Set: Can Analyze: Likely TOI
32	Unknown	Unknown	Test Set: Cannot Analyze
33	Unknown	Unknown	Test Set: Cannot Analyze
2292	Unknown	Unknown	Test Set: Cannot Analyze
77	-	-	Training Set
43	-	-	Training Set
5	-	-	Training Set
954	-	-	Training Set

Extreme Don't Dig Threshold
Other Possible Don't Dig Thresholds
Extreme Don't Dig Threshold

Figure 36: A sketch of a ranked anomaly list, altered to allow comparisons between different training and test sets. Training set anomalies have been appended to the bottom of the list, below even the most extreme don't dig threshold.

Training set anomalies were reflected in the recalculated counts of the number of TOI and Non-TOI digs. This led to a uniform shift of the classification performance curve away from the origin, as shown in Figure 37. The shape of the curve was not altered. For example, if Y of the TOI anomalies had been assigned to the training set, then these Y anomalies must always fall below the don't dig threshold, regardless of which don't dig threshold was used. Therefore, the number of TOI digs must always start at Y. Similarly, if X of the Non-TOI anomalies had been assigned to the training set, then these X anomalies must also always fall below the don't dig threshold. In this way, the number of Non-TOI digs must always start at X. Thus, the gap between the origin and the lower left end of the shifted curve represents the training set anomalies. The smaller the gap, the smaller the training set. The more vertically oriented the gap, the more the training set consisted of TOI anomalies, rather than Non-TOI anomalies. All curves adjusted in this manner were now based on the same total number of TOI and Non-TOI anomalies, regardless of how the anomalies were distributed between the training and test sets. This allowed an "apples-to-apples" comparison between all classification analyses.

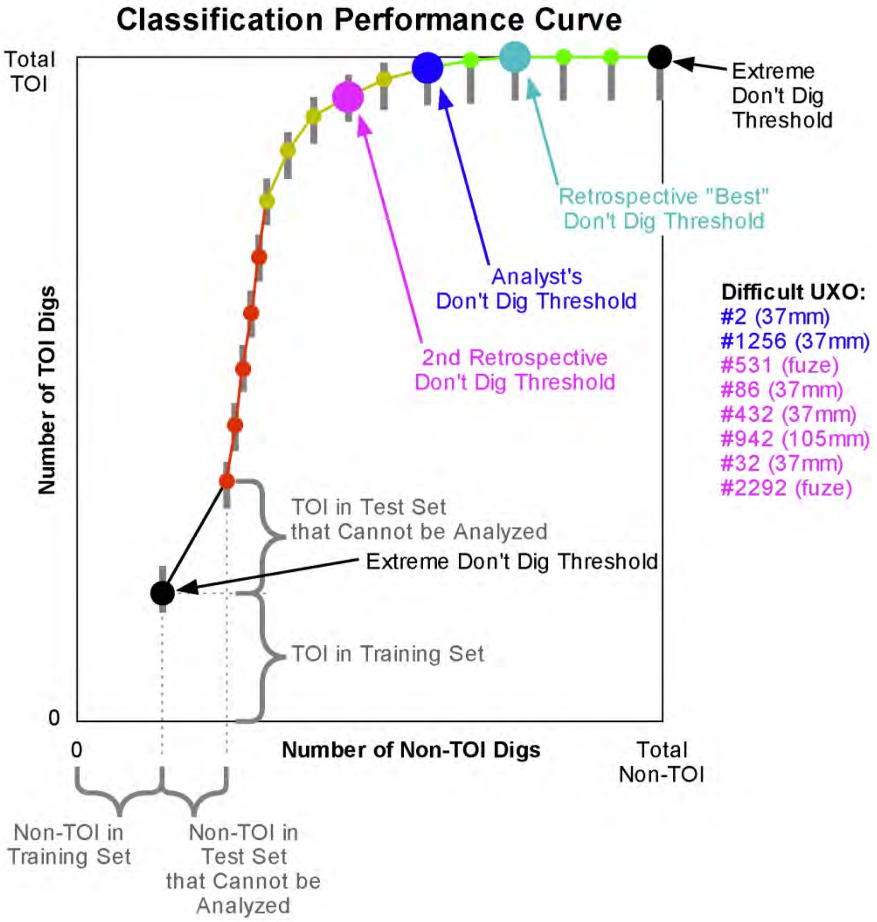

Classification Performance Curve
Total TOI
Extreme Don't Dig Threshold
Retrospective "Best" Don't Dig Threshold
Analyst's Don't Dig Threshold
2nd Retrospective Don't Dig Threshold
Difficult UXO:
#2 (37mm)
#1256 (37mm)
#531 (fuze)
#86 (37mm)
#432 (37mm)
#942 (105mm)
#32 (37mm)
#2292 (fuze)
Number of TOI Digs
TOI in Test Set that Cannot be Analyzed
Extreme Don't Dig Threshold
TOI in Training Set
0
0
Number of Non-TOI Digs
Total Non-TOI
Non-TOI in Training Set
Non-TOI in Test Set that Cannot be Analyzed

Figure 37: A sketch of a classification performance curve, where the training set anomalies have been included in scoring. The curve has been shifted away from the origin, its shape left intact. The gap between the origin and the lower left end of the curve represents the training set anomalies. The 5% most difficult TOIs are listed to the side. TOIs listed in blue and pink fell above and below the analyst's don't dig threshold, respectively.

Different classification analyses were challenged by different anomalies. Figure 37 also lists those TOI anomalies that were the most difficult to classify. Specifically, of the 171 TOI anomalies at the former Camp Butner, the list included the 8 (5%) TOI anomalies listed furthest up the ranked anomaly list. These TOI anomalies had the highest estimated likelihood of being Non-TOIs (i.e., they were classified "the most wrong"). By definition, all eight of these TOI anomalies rose above the second retrospectively chosen don't dig threshold, marked with a large pink dot on the classification performance curve (i.e., the minimum FP while Pd ≥ 95%). Some of these eight TOI anomalies also incorrectly rose above the analyst's prospectively chosen don't dig threshold, marked with a large dark blue dot on the curve; these TOI anomalies were listed in blue. The TOI anomalies listed in pink are those that fell below the dark blue dot but above the pink dot.

ESTCP considered the shapes of the classification performance curves when assessing the performance of the classification analyses. In general classification problems, the area under the curve (AUC) is used to quantitatively describe the shape of the classification performance curve [30]. A curve with a sharp angle near the upper left corner of the plot has a large AUC. In UXO remediation, this would indicate that most Non-TOI anomalies were arranged higher on the ranked anomaly list than most TOI anomalies. That is, the classification algorithm correctly classified most Non-TOI anomalies and most TOI anomalies because the estimated target characteristics overlapped little in multidimensional feature space. However, in UXO remediation, a classification performance curve can indicate good or even excellent performance without a large AUC, since a TOI mistakenly left in the ground (an FN) is considered much worse than a Non-TOI unnecessarily recovered (an FP).

Figure 38 show sketches of two classification performance curves. The left sketch exhibits a large AUC, with a very sharp angle near the upper left corner. Even in retrospect, however, no don't dig threshold could have been chosen to achieve a small number of Non-TOI digs while all TOIs were recovered. In contrast, the right sketch exhibits a smaller AUC, but a don't dig threshold could have been chosen to *halve* the number of Non-TOI digs while still recovering all TOIs. As the DSB pointed out in its 2003 report, even a small reduction in the number of Non-TOI digs can lead to a substantial reduction in the costs of UXO remediation [24]. Thus, although a large AUC is evidence of an algorithm's ability to accurately discriminate between TOIs and Non-TOIs, the true test of an algorithm's performance *in UXO remediation* is to reduce the number of Non-TOI digs while still recovering all TOIs.

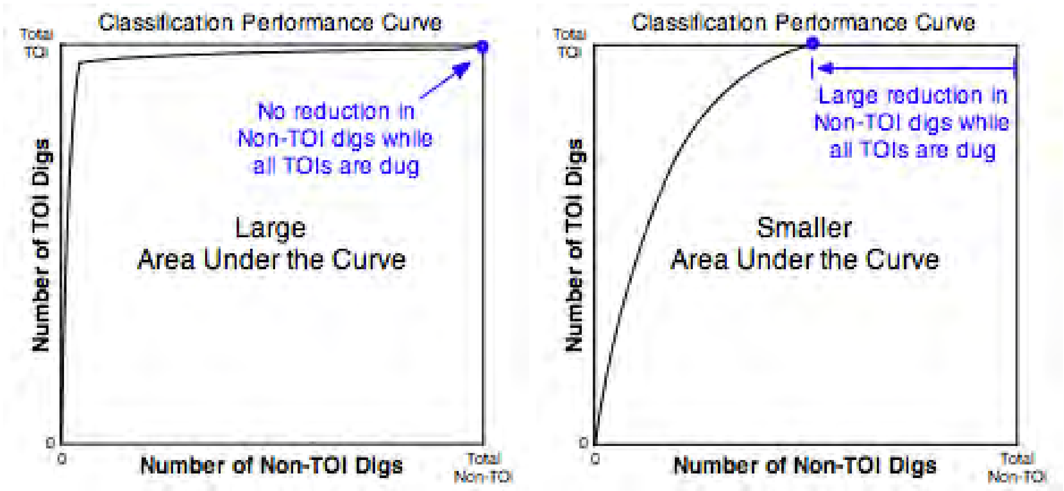


Figure 38: Sketches of two classification performance curves. Left: The large area under the curve indicates a strong ability to discriminate between TOIs and Non-TOIs, but no don't dig threshold could have been chosen to reduce the number of Non-TOI digs while still recovering all TOIs. Right: The smaller area under the curve indicates a weaker ability to discriminate between TOIs and Non-TOIs, but a don't dig threshold could have been chosen to halve the number of Non-TOI digs while still recovering all TOIs. This curve more closely addresses the needs of the UXO community.

In this demonstration, ranked anomaly lists were ultimately judged as follows. First, ESTCP assessed the placement of the analyst's don't dig threshold (dark blue dot) by considering the resulting number of TOI and Non-TOI digs (TP and FP). Ranked anomaly lists with near-maximal TPs and low FPs were desired. As stated in the ESTCP Demonstration Plan [2], a successful classification analysis was defined as one where, at the analyst's don't dig threshold, the number of Non-TOI digs was reduced by at least 30% while all TOIs were correctly classified. Second, ESTCP then assessed what would have been the "best" don't dig threshold (light blue dot). To do this, the minimum number of Non-TOI digs was considered when TP was held at its maximum possible value. This second assessment was made because an otherwise excellent classification analysis (data collection, target characterization, algorithm optimization, etc.) could be marred simply by an inappropriate choice of don't dig threshold. That is, a classification analysis could also be deemed successful if, *in retrospect*, a don't dig threshold could been have chosen to reduce the number of Non-TOI digs by at least 30% while correctly classifying all TOIs [2].

7. Classification Performance Results

This section describes the main results of the UXO classification demonstration at the former Camp Butner. Results are first presented for the EM61-Mk2 cart, since this is currently the standard EMI instrument used for UXO remediation. Results for the TEMTADS and the MetalMapper are presented next; these advanced instruments are not yet well known to the UXO community. Comparisons are made between the different methods used to characterize and classify the buried targets, including the use of commercial versus custom-built software and standard dipole versus more advanced geophysical models. Each comparison is illustrated by a representative set of classification performance curves. A full set of curves, 1 for each of the 54 ranked anomaly lists, can be found in Appendix A. Appendix B lists the classification metrics calculated for each ranked anomaly list at the analyst's don't dig threshold and the retrospective "best" don't dig threshold. Finally, an accompanying DVD contains digital copies of all metrics and curves.

A. Traditional Instrument: EM61-Mk2 Cart

Classification based on the EM61-Mk2 dynamic data exhibited consistently poor performance. Due to the limitations of the EM61-Mk2 cart, size is the only target characteristic that can be accurately estimated from the EM61-Mk2 data. Shape and limited information about material composition and wall thickness can also be estimated, although typically with less accuracy than size. Unfortunately, size was not a strong discriminating feature at the former Camp Butner because many of the TOIs (e.g., 37 mm projectiles and M48 fuzes) were approximately the same size as Non-TOIs (e.g., scrap metal from exploded 105 and 155 mm rounds).

1. No Model

NAEVA, a commercial geophysics company, collected and analyzed the EM61-Mk2 dynamic data. This team's analysis was the most simple and straightforward. Unlike the other classification analyst teams, NAEVA did not invert the EM61-Mk2 data to estimate the polarizabilities of the buried target. Instead, the analyst simply took measurements of the data itself. To do this, she used functions built in to UX-Detect and UX-Process, modules of the Oasis montaj commercial software package sold by Geosoft [5][18][25][26][47][48].

In one analysis, the NAEVA analyst measured the decay rate of the peak amplitude of each detected anomaly [25]. (In contrast, other teams performed geophysical inversions to estimate the decay rates of the principal polarizabilities of each buried target). NAEVA's measurement was a rough estimate of the target's material composition and wall thickness. Slower decay rates were intended to indicate thicker walled, ferrous targets, such as TOIs, while faster decay rates were intended to indicate thinner walled, nonferrous targets, which Non-TOIs often are.

Results were poor. As shown in Figure 39, 6 TOIs were misclassified at the analyst's don't dig threshold (dark blue dot): the number of TOI digs was 165, 6 short of 171, the total number of TOIs. Furthermore, even at the retrospective "best" don't dig threshold (light blue dot), the number of Non-TOI digs could be reduced by only 303 (a reduction of only 14%, from 2120 to 1817) while still digging all TOIs. These results were not unexpected, for two reasons. First, the decay rates were calculated from the data themselves, rather than from polarizabilities inverted from the data using a geophysical model. Second, the short decay interval covered by the time gates of the EM61-Mk2 is often insufficient for classification, regardless of how the decay is calculated (direct measurements versus geophysical inversions). The differences in material composition and wall thicknesses between TOIs and Non-TOIs often do not become apparent until several milliseconds or even tens of milliseconds after the primary field is turned off. This often does not occur until after the latest time gate of the EM61-Mk2.

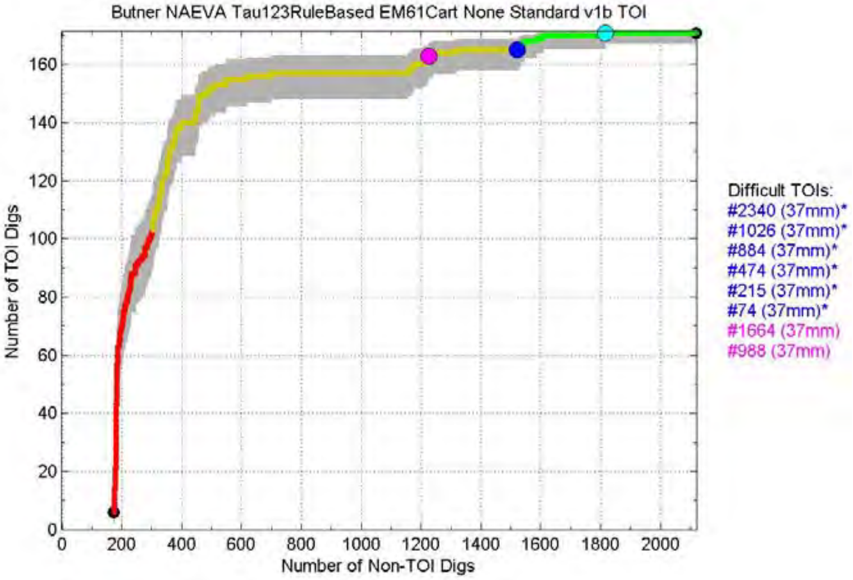


Figure 39: NAEVA's performance based on the decay rates of the peak amplitudes of the anomalies, measured from the EM61-Mk2 dynamic data with UX-Detect and UX-Process and classified with rules optimized over the standard training set.

NAEVA also used an alternative method for classification. With UX-Detect and UX-Process, the analyst measured the peak amplitude of each detected anomaly at the second time gate [5]. This measurement was meant to be an estimate of the size of the buried target. Signal amplitude is only a very rough estimate of size, however, as it can be confounded by target depth. That is to say, a large, deep target can produce the same signal amplitude as a small, shallow target. Recognizing this limitation, the analyst was unable to set a don't dig threshold. Instead, she simply classified all anomalies as "Cannot Decide." This is equivalent to setting the don't dig threshold at the very top of the ranked anomaly list, such that all anomalies fall below threshold. As a result, all points of the classification performance curve in Figure 40 are colored in yellow, and the dark blue dot rests at the upper right end of the curve.

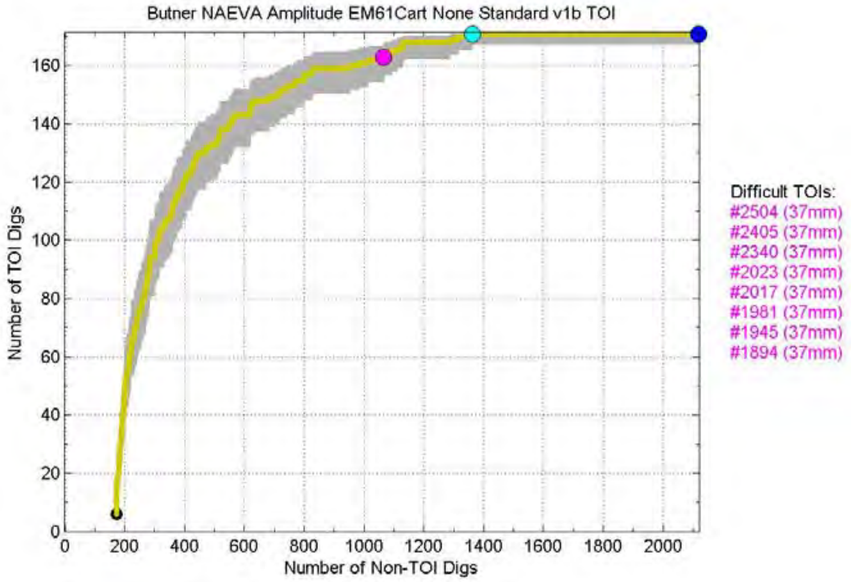


Figure 40: NAEVA's performance based on the peak amplitudes of the anomalies, measured from the EM61-Mk2 dynamic data with UX-Detect and UX-Process and classified with rules optimized over the standard training set. All anomalies in the standard test set were declared "Can't Decide" (yellow), as the don't dig threshold could not be set prospectively.

The retrospective results were better than expected, however. The number of Non-TOI digs could have been reduced by 756 (a reduction of 36%, from 2120 to 1364) while still digging all TOIs. This result is illustrated by the light blue dot in Figure 40. Surprisingly, peak signal amplitude, although confounded by depth, still proved to have some discriminating power at the former Camp Butner. This result has not been seen in other studies, however. Furthermore, these results can only be considered favorable in a retrospective sense; due to the uncertainties inherent in using signal amplitude as the sole feature for classification, the NAEVA analyst could not prospectively set the don't dig threshold and therefore would not have been able to use this method to reduce the number of unnecessary digs in a real remediation project.

2. Dipole Model

Other teams used geophysical models to invert the EM61-Mk2 data. SAIC used the commercially available UX-Analyze software to perform the inversions while Sky used the UXOLab software developed by the University of British Columbia [1][22][38][39][50][51]. Both teams estimated the three polarizabilities of the buried target and then calculated features related to the target's size and material composition/wall thickness. While their size feature was based on the sum of the three polarizabilities, their material composition/wall thickness feature was based on the decay rate of the principal polarizability. Neither SAIC nor Sky calculated a feature related to the shape of the buried target for their EM61-Mk2 analyses. An estimate of shape would have involved the ratio of one polarizability to another, requiring accurate estimates of each individual polarizability. This is difficult to reliably achieve with the EM61-Mk2 dynamic data because the cart's line-to-line position uncertainty makes it difficult to constrain the individual polarizabilities during the inversion process. Instead, only the sum of the polarizabilities can be well constrained (an indication of size). This sum is the trace of the diagonalized polarizability matrix, which is tensor invariant.

Results were poor for both teams. Figure 41 shows that at SAIC's don't dig threshold, 1 TOI was misclassified while the number of Non-TOI digs could be reduced by only 319 (a reduction of only 15%, from 2120 to 1801). This reduction was even smaller at the retrospective "best" don't dig threshold. Figure 42 shows similar results for Sky. Here, 2 TOIs were misclassified at Sky's don't dig threshold; the number of Non-TOI digs could be reduced by 510 (a reduction of 24%, from 2120 to 1610). As with SAIC, Sky's reduction in Non-TOI digs was much smaller at the retrospective "best" don't dig threshold.

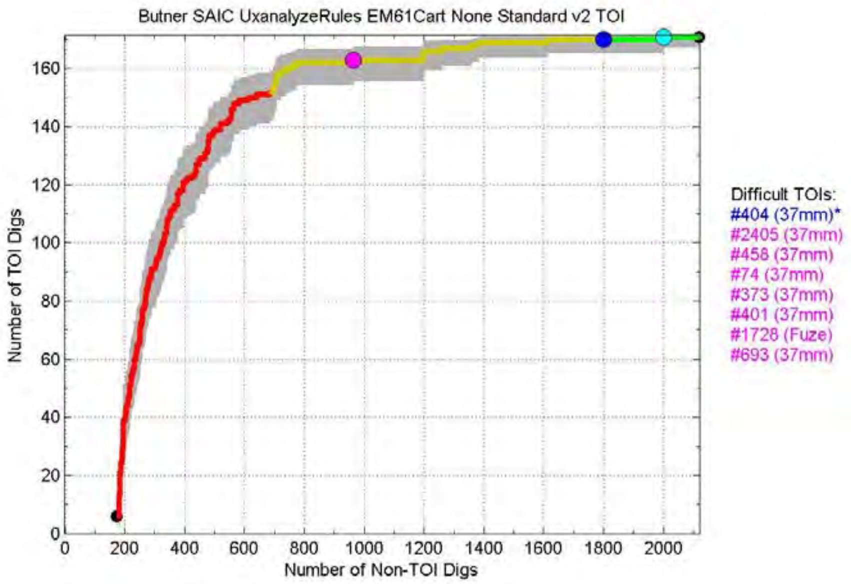


Figure 41: SAIC's performance for a retrospective analysis based on amplitudes and decay rates of polarizabilities inverted from the EM61-Mk2 dynamic data using the UX-Analyze/IDL multi-source dipole model and classified with a rules-based library matcher optimized over the standard training set.

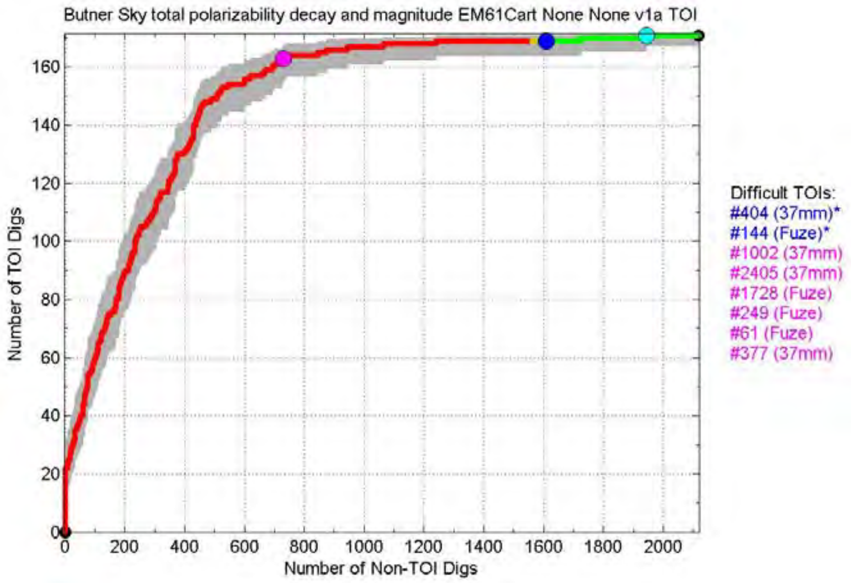


Figure 42: Sky's performance based on amplitudes and decay rates of polarizabilities inverted from the EM61-Mk2 dynamic data using the UXOLab single-source and two-source dipole models and classified with a statistical classifier optimized over an existing library.

B. Advanced Instruments: TEMTADS and MetalMapper

Classification based on the TEMTADS and MetalMapper data exhibited good and sometimes excellent results. The MetalMapper in dynamic mode shares many of the same limitations as the EM61-Mk2 cart; size therefore is the main target characteristic that can be estimated from the MetalMapper dynamic data. (Limited information about shape and material composition/wall thickness can also be estimated, although not as accurately as size). Neither the TEMTADS nor the MetalMapper in static mode is limited to such a degree. The target's size, as well as its shape and material composition/wall thickness, can be accurately estimated from these static data. This led to good and sometimes excellent results because shape and material composition/wall thickness were strong discriminating features at the former Camp Butner (size was not). All analyses of the TEMTADS and MetalMapper data used geophysical models for inversion. The models varied in complexity. The analyses also varied in other ways, such as in the algorithms used for classification and the training sets used to optimize the algorithms.

1. Dipole Model

Sky was the only team to analyze the MetalMapper dynamic data. This team used UXOLab to invert the data, estimating the polarizabilities of each buried target. Analysts selected a subset of features related to the targets' sizes and material compositions/wall thicknesses and then used a statistical classifier to classify the targets based on the selected subset of features [22].

Results were better than those based on the EM61-Mk2 dynamic data. Figure 42 showed the classification performance curve for Sky's EM61-Mk2 analysis, and Figure 43 now shows the corresponding curve for Sky's dynamic MetalMapper analysis. The analyst's don't dig threshold was poorly set for the MetalMapper curve, resulting in the misclassification of 19 TOIs. However, the MetalMapper curve exhibited a better shape than the EM61-Mk2 curve. In fact, at the MetalMapper curve's retrospective "best" don't dig threshold, the number of Non-TOI digs could have been reduced by 717 (a reduction of 34%, from 2120 to 1403). This is a much larger reduction than what was shown in the EM61-Mk2 curve. This shows, at least in the retrospective sense, that aside from the selection of the don't dig threshold, all other aspects of Sky's MetalMapper classification analysis performed reasonably well.

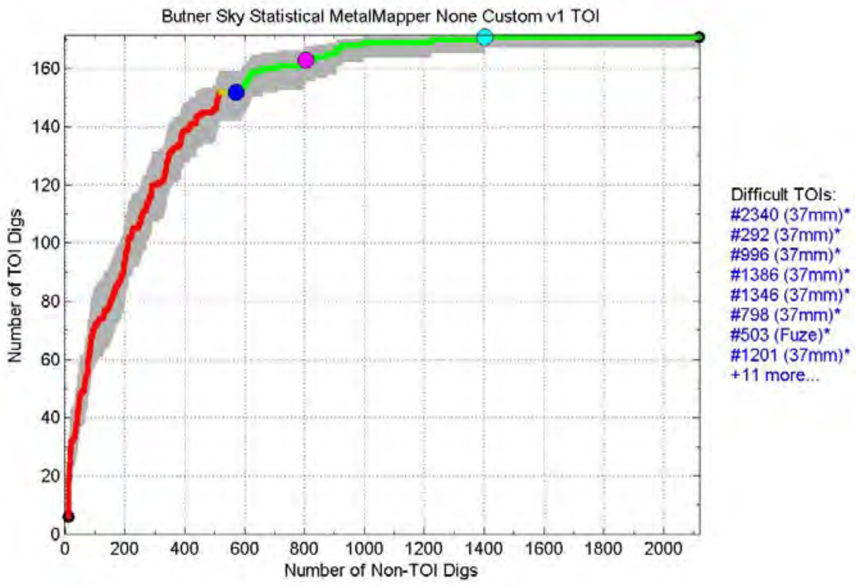


Figure 43: Sky's performance based on amplitudes and decay rates of polarizabilities inverted from the MetalMapper dynamic data using the UXOLab single-source and two-source dipole models and classified with a statistical classifier optimized over a custom training set.

The MetalMapper in dynamic mode exhibited a better position accuracy than the EM61-Mk2 cart due to its seven triaxial receive coils. The individual polarizabilities were better constrained during inversion, resulting in more accurate estimates of the targets' sizes. Useful polarizability decay rates were more difficult to estimate, however. Because the MetalMapper in dynamic mode was configured to use an even a shorter time interval than the EM61-Mk2, the polarizability decays were not assessed for a long enough time to make them very useful for classification. One could attempt to improve the classification performance of the MetalMapper in dynamic mode by extending its time gates; this would lead to better calculation of the polarizability decay rates, which in turn would lead to more accurate estimates of the target's material composition/wall thickness. However, extending the time gates would also slow the data collection procedure. Thus, a trade-off must be made [51].

Sky also analyzed the MetalMapper static data, as did many other classification analyst teams. Sky input the full polarizability curves (containing information about size, shape, material composition, and wall thickness) into a statistical classifier [50][51]. Figure 44 shows Sky's classification performance curve. Four TOIs were misclassified at the analyst's don't dig threshold. All four were 37 mm projectiles, the smallest and therefore most difficult TOIs to classify. The number of Non-TOI digs was reduced by 1983 at the analyst's don't dig threshold, a full 94% (from 2120 to 137). The reduction in Non-TOI digs was much smaller at the retrospective "best" don't dig threshold, however.

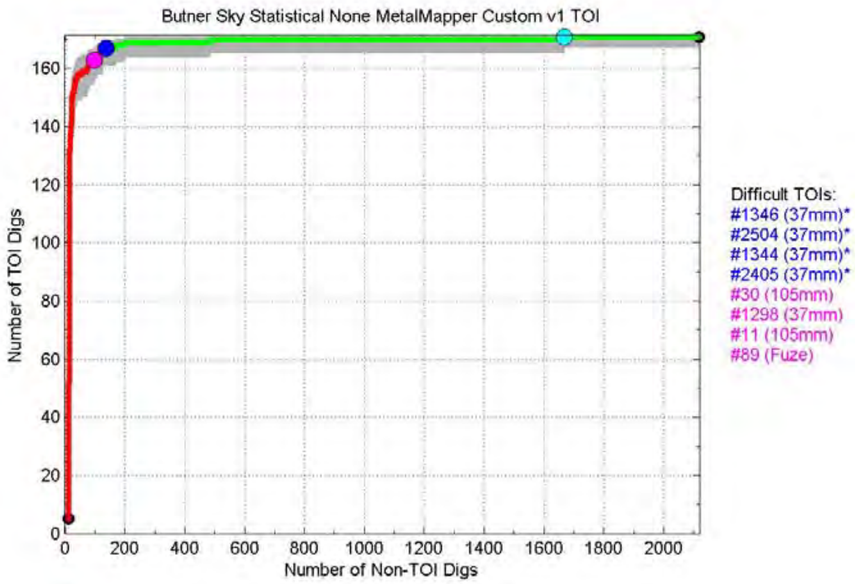


Figure 44: Sky's performance based on full polarizability curves inverted from the MetalMapper static data using the UXOLab single-source and two-source dipole models and classified with a statistical classifier optimized over a custom training set.

The MetalMapper in static mode led to good classification performance because it did not suffer from the same limitations as the dynamic instruments. In static mode, all three transmit coils were used, fully illuminating the target in all three directions. The three individual polarizabilities were therefore well constrained during the inversion process, leading to very accurate estimates. Furthermore, successive measurements of the received signal were stacked, increasing the SNR of the received signal and further increasing the accuracy of the estimated polarizabilities. Finally, the time gates in static mode were set much longer than in dynamic mode. This provided a longer expanse of time over which the polarizabilities' decay rates could be assessed, allowing the differences between TOIs and Non-TOIs to become more evident.

Finally, Sky analyzed the TEMTADS static data, as did many other classification analyst teams. Once again, Sky input the full polarizability curves into a statistical classifier [50][51]. This analysis led to excellent performance, as illustrated by the curve in Figure 45. The analyst's don't dig threshold was appropriately set, leading to a near maximal reduction in Non-TOI digs (1941, a reduction of 92% from 2120 to 179) while correctly classifying all TOIs. The TEMTADS shares many of the same advantages of the MetalMapper in static mode, including the ability to illuminate the target in all three directions and stacking to improve SNR. In addition, the excellent performance of the TEMTADS is likely due to its very long time gates (25 ms), leading to material composition/wall thickness estimates with very fine resolution. Its improvement in performance over the MetalMapper in static mode may also be due to the careful field technique exhibited by the TEMTADS data-collection team.

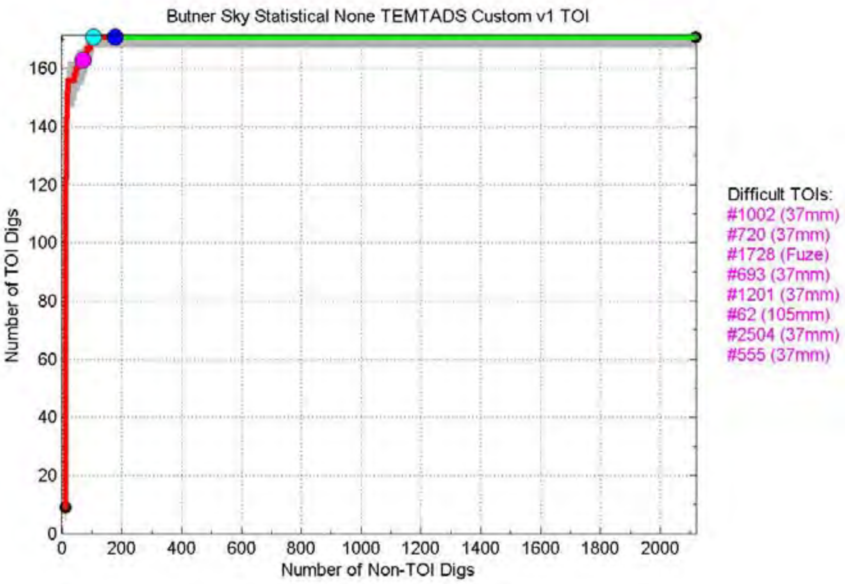


Figure 45: Sky's performance based on full polarizability curves inverted from the TEMTADS static data using the UXOLab single-source and two-source dipole models and classified with a statistical classifier optimized over a custom training set.

a. Differences in MetalMapper Systems

Some classification analyst teams commented on the differences in SNR between the MetalMapper static data collected by Geometrics and those collected by Sky [47][50]. Sky used a newer MetalMapper system, which used better cabling, was positioned lower to the ground, and was pulled by a tractor with an electromagnetically quieter engine. As part of the failure analysis after the end of the demonstration, Sky investigated how these differences affected classification performance. They performed one classification analysis using only those data collected by the newer Sky system. Then they independently performed a second analysis using only those data collected by the older Geometrics system [50].

Figure 46 shows the classification performance curves resulting from the two analyses. Sky plotted these curves using slightly different metrics than the other curves presented this far. The percentage of TOI dug (i.e., Pd) is plotted on the vertical axis while the horizontal axis plots the percentage of Non-TOI dug (i.e., Pfa, but called FAR in this analysis). The newer Sky system (red curve) outperformed the older Geometrics system (black curve). At the retrospective "best" don't dig threshold (arrows), the Sky system could have reduced the number of Non-TOI digs by 99%, leading to a FAR of only 1%. In contrast, the Geometrics system could have reduced this number by only 63%, leading to a FAR of 37%. Both Sky and NAEVA recommended using more conservative classification approaches for data with low SNRs, such as the data collected with the older MetalMapper system. For Sky, this meant choosing a different subset of target characteristics on which classification would be based, that is, size and material composition/wall thickness features only, rather than the full polarizability curves [50]. For NAEVA, this meant relaxing some rules for classifying a target as “Likely TOI” [47].

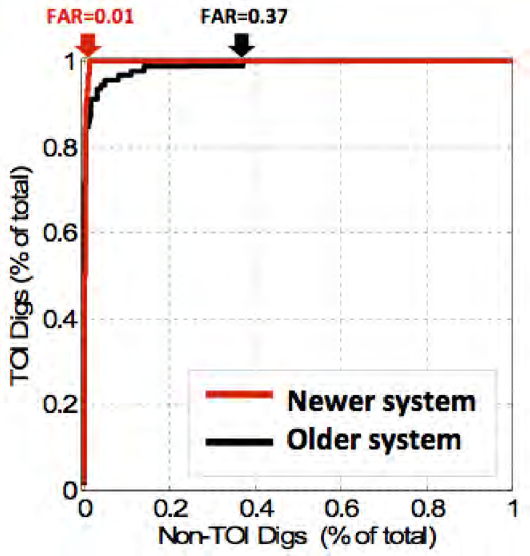


Figure 46: Sky’s performance based on data collected from its newer MetalMapper system (red) and Geometrics’ older system (black). The percentage of TOIs dug (i.e., Pd) is plotted versus the percentage of Non-TOIs dug (i.e., Pfa or FAR). Arrows mark the retrospective “best” don’t dig thresholds. Taken from [50].

b. Static Data Requests

Some classification analyst teams requested static data for only some anomalies, those for which the EM61-Mk2 dynamic data were deemed inadequate for analysis. Other teams requested no static data, relying solely upon the EM61-Mk2 dynamic data. Still other teams requested static data for all anomalies. SAIC and Sky performed all three types of analyses [1][22]. Figure 41 has already illustrated the poor performance of SAIC’s analysis of the EM61-Mk2 dynamic data. Figure 47 and Figure 48 now show the performance of SAIC’s analyses of TEMTADS static data for some and all anomalies, respectively. The analysis based on only *some* static data (Figure 47) also exhibited poor performance, as 10 TOIs were misclassified at the analyst’s don’t dig threshold, including at least 1 M48 fuze. Furthermore, even retrospectively, the don’t dig threshold could not be adjusted to reduce the number of Non-TOI digs while correctly classifying all TOIs, as evidenced by the light blue dot resting at the upper right end of the curve. As part of its failure analysis, SAIC discovered that some TOIs had been misclassified based solely on the EM61-Mk2 dynamic data, before the static data were even requested. Had the static data been available, these TOIs would not have been misclassified [38]. In fact, the analysis based on *all* static data (Figure 48) exhibited much better classification performance—as only three TOIs were misclassified at the analyst’s don’t dig threshold. Furthermore, the retrospective “best” don’t dig threshold could have reduced the number of Non-TOI digs by 433 (a reduction of 20%, from 2120 to 1687) while correctly classifying all TOIs.

These results were disappointing. This demonstration was specifically designed so that analysts could choose which anomalies required static data, in the hopes that fewer static data would be needed to produce the same classification performance. Unfortunately, this did not turn out to be the case. The analyses based on only some static data resulted in a much poorer performance, even though the amount of static data was reduced only minimally. For example, SAIC requested TEMTADS static data for 86% of the anomalies. In a real remediation project, collection of static data for the remaining 14% would not have greatly increased the cost of the project—but would have greatly improved performance. Other teams also requested large amounts of static data—Sky and Parsons requested static data for 78% and 69% of the anomalies, respectively. Neither of these analyses exhibited particularly good performance either.

c. Training Data Requests

Some classification analyst teams requested custom training sets, based on those anomalies in the demonstration area that they deemed most likely to optimize their classification algorithms. In general, the custom training sets were smaller and consisted of a larger percentage of TOIs than the Standard Training Set did. For example, Figure 39 and Figure 40 showed classification performance curves for two of NAEVA's analyses based on the EM61-Mk2 dynamic data. In each case, the gap between the origin and the lower left end of the curve represents the training set. The gap is larger in size and more horizontally oriented than the other curves shown so far. This occurs because NAEVA used the Standard Training Set to optimize its algorithms; the other curves were based on either a custom training set (leading to a smaller, less horizontally oriented gap) or existing data from the IVS, training pit, and previous studies only (leading to no gap at all). Specifically, the Standard Training Set consisted of 179 anomalies (3% of which were TOIs); the custom training sets (mean and standard deviation) consisted of only 108 ± 71 anomalies (20% ± 12% of which were TOIs). In a real remediation project, a smaller training set would lead to lower costs because fewer targets would have to be recovered to provide ground truth for optimizing the classification algorithms.

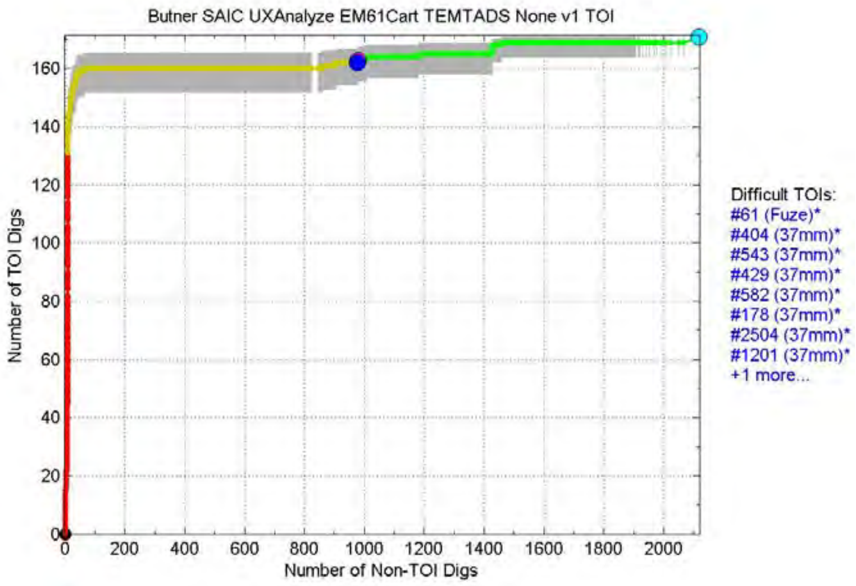


Figure 47: SAIC's performance based on amplitudes, ratios, and decay rates of polarizabilities inverted from the EM61-Mk2 cart dynamic data and requested TEMTADS static data using the UX-Analyze/IDL multi-source dipole model and classified with a rules-based library matcher optimized over an existing library.

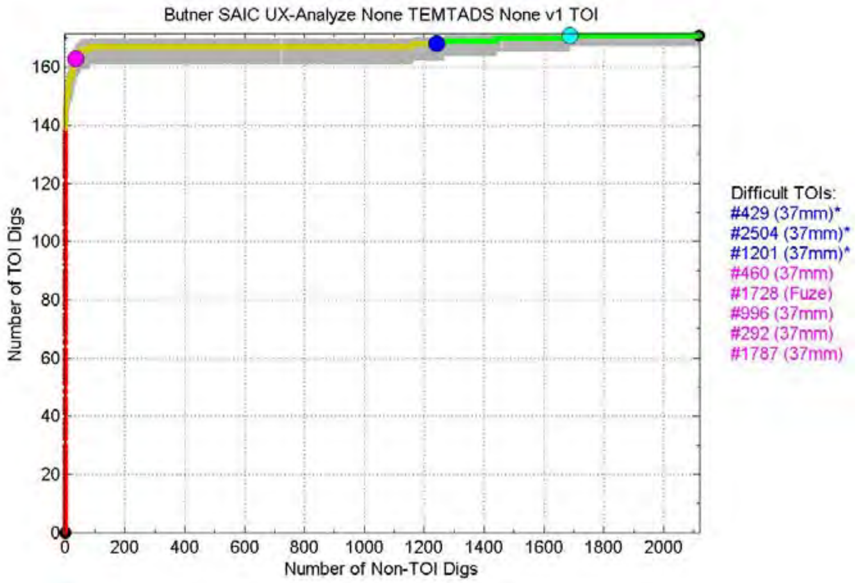


Figure 48: SAIC's performance based on amplitudes, ratios, and decay rates of polarizabilities inverted from the TEMTADS static data using the UX-Analyze/IDL multi-source dipole model and classified using a rules-based library matcher optimized over an existing library.

UXO classifications algorithms can often be satisfactorily optimized using only a very small training set because the classification is based on quantifying the boundaries around clusters of TOIs, rather than Non-TOIs, in multidimensional feature space. Then, all test set anomalies that fall within these boundaries can be classified as “Likely TOI,” and all that fall outside can be classified as “Likely Non-TOI.” Analysts do not have to use data and ground truth collected from the demonstration area to set these boundaries. Instead, they can simply use the data and ground truth collected from the TOIs seeded in the IVS and training pit. They can also use data and ground truth collected from TOIs in previous analyses at other sites.

There is one caveat, however: Some TOIs in the demonstration area may turn out to be of an unexpected munition type. As a result, TOIs of that type would not have been seeded in the IVS and training pit, and the analyst would not have known to use data and ground truth collected in previous studies from that TOI type. Therefore, there is a risk that the analyst may not realize that a cluster in multidimensional feature space represents TOIs. To mitigate this risk, some analysts choose to use a small custom training set consisting of a subset of anomalies in the demonstration area. These anomalies are carefully chosen to sample every cluster in multidimensional feature space. Once the anomalies are recovered and their ground truth is available as training data, the analyst can then properly identify which clusters represent TOIs and which represent Non-TOIs and therefore more appropriately set the boundaries encircling the TOI clusters.

For example, historical records showed that 37 mm projectiles had been previously fired at the former Camp Butner [19]. Therefore, ESTCP seeded one 37 mm projectile in both the IVS and training pit, as well as 110 in the demonstration area. Data and ground truth for the IVS and training pit seeds were made available to the classification analyst teams for their algorithm optimization. Some targets in the demonstration area exhibited polarizabilities similar to those estimated from the 37 mm projectiles seeded in the IVS and training pit. Therefore, many analysts correctly concluded that these demonstration area targets were 37 mm projectiles as well. For example, Figure 49(a) shows the polarizabilities estimated from three targets in the demonstration area (black, pink, and light blue); these polarizabilities were very similar to those estimated from a 37 mm projectile obtained from the U.S. Army Corps of Engineers (dark blue), as well as the 37 mm projectiles seeded in the IVS and training pit. Figure 49(b) shows photographs of these three demonstration area targets; indeed, all turned out to be 37 mm projectiles. In fact, all turned out to have driving bands, just as those seeded in the IVS and training pit. On the other hand, some targets in the demonstration area exhibited polarizabilities that were different from those in the IVS and training pit, such as those shown in Figure 49(c) (dashed red, blue, and green). Targets like these formed a new “mystery” cluster in multidimensional feature space. Some classification analyst teams chose to include some of these targets in their custom training sets, so that their ground-truth labels could be learned and the new cluster could be better understood. Figure 49(d) shows photographs of these three targets, all of which turned out to be 37 mm projectiles *without* driving bands. In addition, a third type of 37 mm projectile was also found to be native to the site.

These native TOIs had driving bands but were of a slightly different size and shape than their seeded counterparts. Thus their estimated polarizabilities were also slightly different, forming a third cluster in multidimensional feature space.

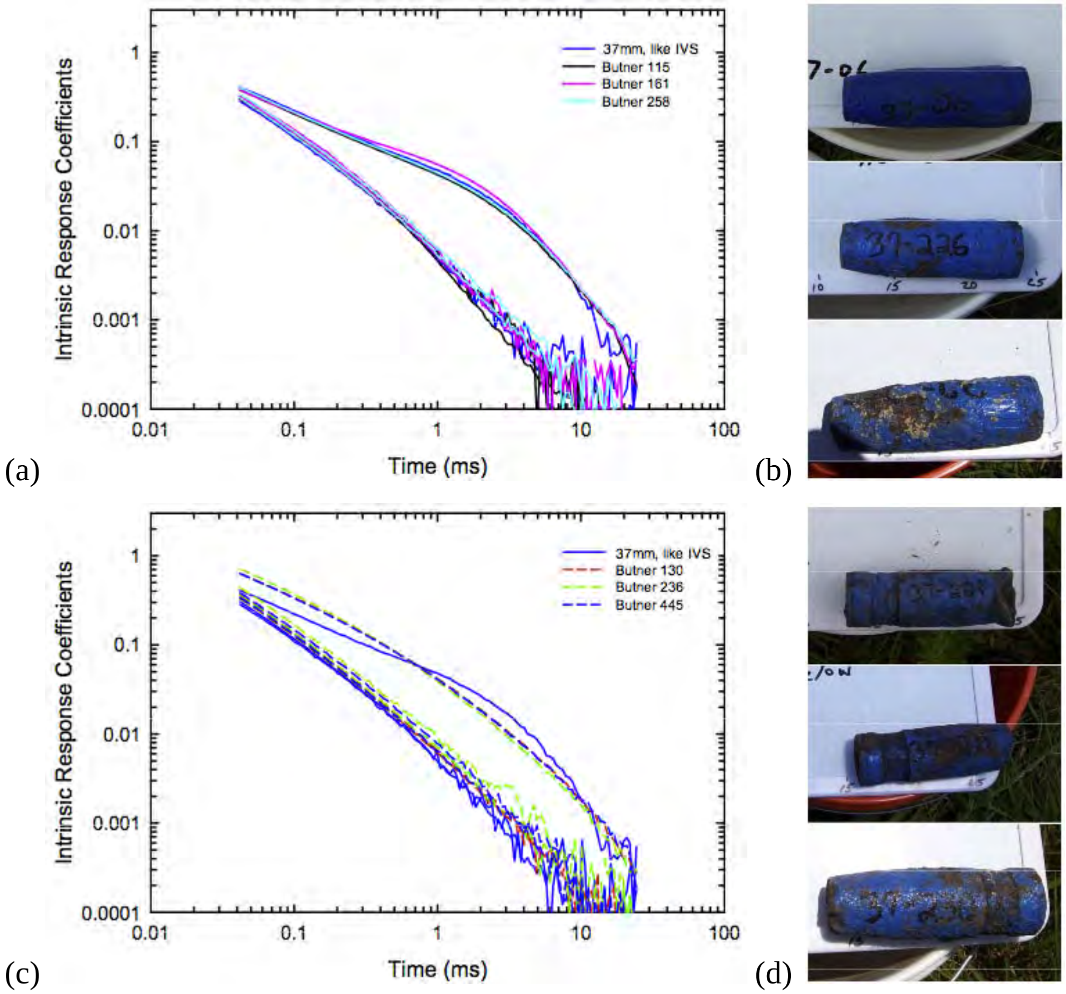


Figure 49: The effect of driving bands on target characterization. (a) Polarizabilities estimated from three targets in the demonstration area (black, pink, and light blue), all similar to the polarizabilities estimated from a 37 mm projectile like those seeded in the IVS and training pit (dark blue). (b) Photographs of the three demonstration area targets, all 37 mm projectiles with driving bands like those seeded in the IVS and training pit. (c) Polarizabilities estimated from three other targets in the demonstration area (dashed red, green, and dark blue), all similar to each other but different from the polarizabilities estimated from a 37 mm projectile like those seeded in the IVS and training pit (dark blue). (d) Photographs of the three other demonstration area targets, all 37 mm projectiles *without* driving bands. Taken from [4].

d. Second-Pass Analyses

SAIC and Sky experimented with re-optimizing their classification algorithms in what became known as "second-pass" analyses. In the first pass, the analyst optimized his classification algorithm using a small training set. The analyst created the ranked anomaly list and submitted it for scoring. He then requested the ground truth for all anomalies classified as "Likely TOI," which, by definition, fell below the don't dig threshold. This mimicked what could occur in a real remediation project, where all anomalies that fell below the don't dig threshold would be dug, and the resulting ground truth could be made available to the classification analyst. The analyst could then use this additional ground truth to assess how well his classification methods had performed so far. If necessary, revisions could be made to those anomalies that had not yet been dug.

In the second pass, then, the analyst re-optimized his classification algorithm based on the additional ground truth of those anomalies that had already been "dug." He then revised his ranked anomaly list based on the newly re-optimized classification algorithm. Some restrictions were imposed, however, to mimic the logical flow of a real remediation project. The analyst could re-rank and reorder only those anomalies that had risen above the first-pass don't dig threshold. All other anomalies must remain unchanged—in a real remediation project, they would have already been dug by this point. Similarly, the analyst could move the don't dig threshold up, but not down, the ranked anomaly list. In a real remediation project, stakeholders could decide to dig more anomalies that had not yet been dug, but could not decide to un-dig anomalies that had already been dug.

Results were promising. Figure 50 and Figure 51 show the classification performance curves for SAIC's first- and second-pass analyses of MetalMapper static data. The first-pass results are poor: 13 TOIs were misclassified at the analyst's don't dig threshold and the retrospective "best" don't dig threshold could reduce the number of Non-TOI digs by only 447 (a reduction of 21%, from 2120 to 1673) while correctly classifying all TOIs. The effects of the second pass of analysis are clearly evident in the second curve. First, the training set is much larger, as evidenced by the very large gap between the origin and the lower left end of the curve. This was expected, as the second-pass training set consisted of all anomalies classified as "Likely TOI" in the first-pass analysis. Second, the curve more closely tracks the upper edge of the plot because some of the TOI anomalies that had initially been misclassified as "Likely Non-TOI" were now more accurately reclassified as "Cannot Decide" or "Likely TOI." Finally, the analyst's don't dig threshold was moved in the upper right direction along the curve, as the analyst had moved the threshold up the ranked anomaly list. These effects improved results. At the analyst's revised don't dig threshold, only 7 TOI were misclassified, as opposed to the 13 that had been misclassified in the first-pass analysis. Furthermore, at the retrospective "best" don't dig threshold, the number of Non-TOI digs could be reduced by over 150 more than in the first-pass analysis.

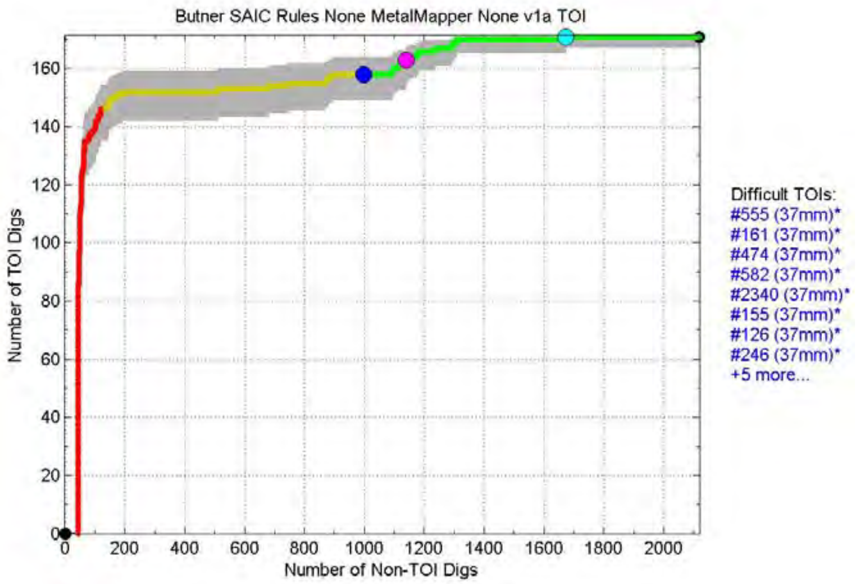


Figure 50: SAIC's performance based on amplitudes, ratios, and decay rates of polarizabilities inverted from the MetalMapper static data using an IDL single-source dipole model and classified with a rules-based library matcher optimized over an existing library.

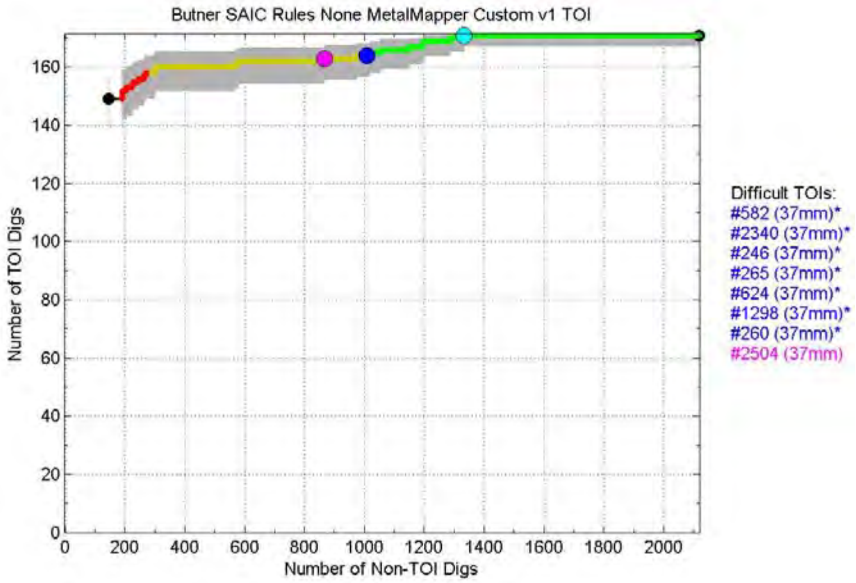


Figure 51: SAIC's performance based on amplitudes, ratios, and decay rates of polarizabilities inverted from the MetalMapper static data using an IDL single-source dipole model and classified with a rules-based library matcher optimized over a custom training set. The custom training set was built from the "Likely TOI" anomalies in the previous figure.

Second-pass analyses, combined with seeded UXO, could give stakeholders more confidence in the final don't dig thresholds used in real remediation projects. That is, classification analysts could submit their first-pass ranked anomaly lists, and stakeholders could then ascertain, *before digging even begins*, how many of the seeded UXO had been correctly classified. For example, the anomalies that SAIC classified as “Likely TOI” in the first-pass analysis included all 26 of the seeded 105 mm projectiles, but only 22 of the 24 seeded M48 fuzes and only 88 of the 110 seeded 37 mm projectiles. Even in the absence of information about native UXO, the stakeholders would already know that the don’t dig threshold was not correctly set because not all the seeded projectiles were correctly classified. At this point, the stakeholders could require the classification analysts to perform a second-pass analysis, and possibly a third- and fourth-pass analysis. That is, this process could iterate until stakeholders had confidence in the don’t dig threshold (e.g., all seeded UXO had been correctly classified). At that point, the digging could finally begin. Of course, in a real remediation project, the number of seeds could be much smaller than in this demonstration, but the principle illustrated above would still hold.

Further passes could also occur, as well. As the anomalies are dug, ground truth could be recovered and made available to the classification analyst. The analyst could further re-optimize his classification algorithm, as was done in this demonstration. This process could also iterate several times, until the stakeholders decide that no further improvement could be made (e.g., the last several dozen digs resulted in no native UXO, etc.). At this point, remediation could finally be considered complete.

e. Analysts’ Classification Experience

In this demonstration, teams new to classification were mentored by more experienced teams. Sky mentored CH2M HILL, and SAIC mentored NAEVA and Parsons. CH2M HILL, NAEVA, and Parsons are commercial geophysics companies that often perform real survey or remediation projects. Their expertise lies in more hands-on fieldwork. In contrast, Sky and SAIC have been involved with the research and development of classification technologies over the past several years.

SAIC gave NAEVA brief instructions on how to use UX-Analyze to invert and classify MetalMapper static data [47]. (NAEVA had previous experience using UX-Analyze with EM61-Mk2 data). The training proved successful. Figure 50 and Figure 51 showed SAIC's first- and second-pass analyses of the MetalMapper static data. Although the first-pass results were poor, the second-pass results were better (at least at the retrospective “best” don’t dig threshold). NAEVA’s results were even better. Figure 52 now shows results of one of NAEVA’s retrospective analyses of the MetalMapper static data. The curve exhibits a sharp angle, indicating a fairly accurate ranking of the anomalies on the ranked anomaly list. Furthermore, the analyst’s don’t dig threshold is nearly optimally placed, correctly classifying all TOIs while reducing the number of Non-TOI digs by 1115 (a reduction of 53%, from 2120 to 1005), almost the maximum amount possible for this curve.

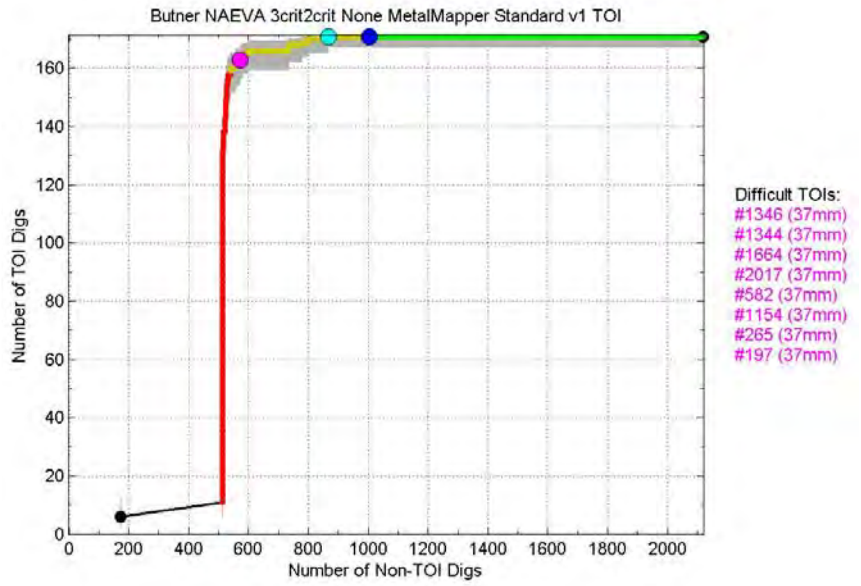


Figure 52: NAEVA's performance for a retrospective analysis based on polarizabilities inverted from the MetalMapper static data using the UX-Analyze single-source dipole model and classified with a rules-based library matcher optimized over the standard training set.

CH2M HILL, who had not previously participated in a live-site demonstration, received more extensive training from Sky. An analyst from CH2M HILL spent time at the Sky offices learning how to invert and classify the MetalMapper static data with the UXOLab software [42]. This training also proved to be successful. Figure 53 and Figure 54 illustrate the performance of Sky's and CH2M HILL's analyses of MetalMapper static data, respectively. The curves are approximately the same shape, indicating that the anomalies were ranked and ordered in approximately the same way. The dark blue dots fall at approximately the same points along the curves, indicating that the analysts' don't dig thresholds were set at approximately the same places on the ranked anomaly lists. In one sense, though, the CH2M HILL curve is actually better than the Sky curve because the retrospective "best" don't dig threshold is better placed. That is, CH2M HILL's analysis could have reduced the number of Non-TOI digs by 1135 (a reduction of 54%, from 2120 to 985) while correctly classifying all TOIs. In contrast, Sky's analysis could have reduced this number by only 14 (a reduction of less than 1%), as evidenced by the light blue dot placed near the upper right end of the curve.

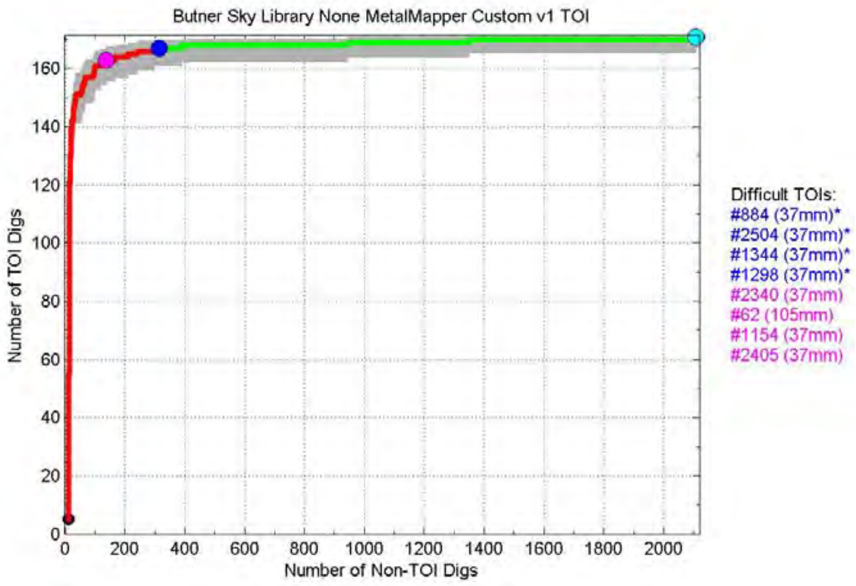


Figure 53: Sky's performance based on full polarizability curves inverted from the MetalMapper static data using the UXOLab single-source and two-source dipole models and classified with a library matcher optimized over a custom training set.

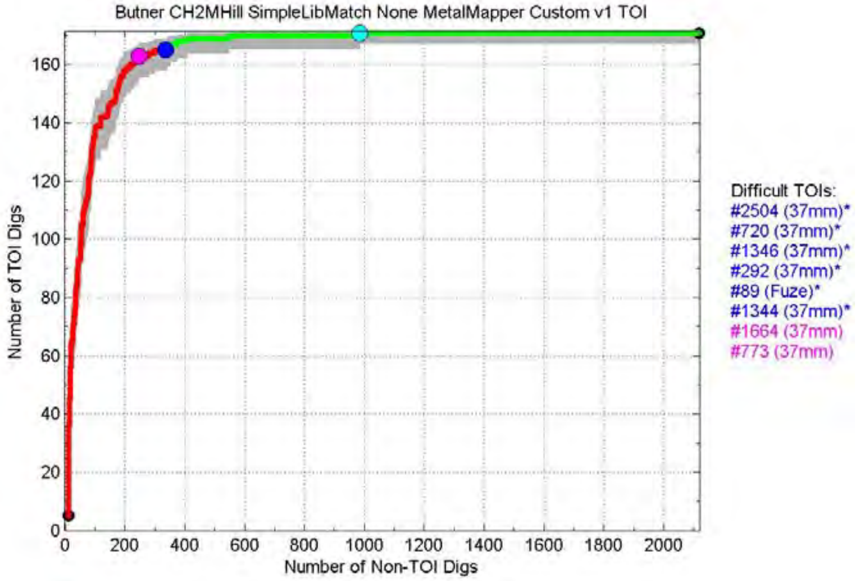


Figure 54: CH2M HILL's performance based on full polarizability curves inverted from the MetalMapper static data using the UXOLab single-source and two-source dipole models and classified with a library matcher optimized over a custom training set.

In both cases, then, the commercial geophysics companies outperformed the more experienced organizations. This may be because NAEVA and CH2M HILL performed fewer different types of analyses, allowing them to focus on, and devote more time and resources to, each individual analysis. In contrast, SAIC and Sky performed several different types of analyses, experimenting with different data sets, classification algorithms, software tools, etc. It may have been difficult for SAIC and Sky to devote the same time and resources to each individual analysis.

Regardless, results show that trained geophysicists with minimal experience in this particular area can perform UXO classification. With some training, analysts from commercial geophysics companies can become familiar with the data collected by the more advanced instruments, such as the MetalMapper, and can learn how to use new software tools, such as the inversion and classification modules embedded within UX-Analyze and UXOLab. That is to say, tech transfer appears very feasible.

f. Classification Algorithms

Some analyst teams experimented with different classification algorithms in an effort to determine which algorithms performed better. In most cases, all algorithms performed similarly under similar conditions. For example, Figure 43 showed the classification performance curve for Sky's analysis of MetalMapper dynamic data when a statistical classifier was used. Sky also experimented with other classification algorithms. Figure 55 now illustrates Sky's performance when a library-matching algorithm was used, and Figure 56 shows the performance of the library-matcher in conjunction with a human expert. All three curves are similar to each other. These results show that when all other factors were equal, the classification algorithm itself did not have a large effect on performance. Of prime importance, then, must have been the other factors of analysis, such as the criteria used to make the Can vs. Cannot Analyze decision, the geophysical inversions, the selection of which features to input into the classification algorithm, and/or the training set on which the algorithm was optimized.

g. Number of Dipole Sources

Most classification analyst teams used dipole models to perform geophysical inversions. While some teams assumed only one dipole source per anomaly, others assumed multiple dipole sources. For example, Figure 48 showed SAIC's performance when analyzing the TEMTADS data with the UX-Analyze software. The software called a custom-built IDL module that used an iterative process to arrive at the number of dipole sources that best fit the data. In contrast, Figure 57 now illustrates a second analysis by SAIC, one based on custom-built software that assumed only one dipole source per anomaly. All other factors in the analysis remained the same, including the Can vs. Cannot Analyze criteria, the classification algorithm, and the training set [37].

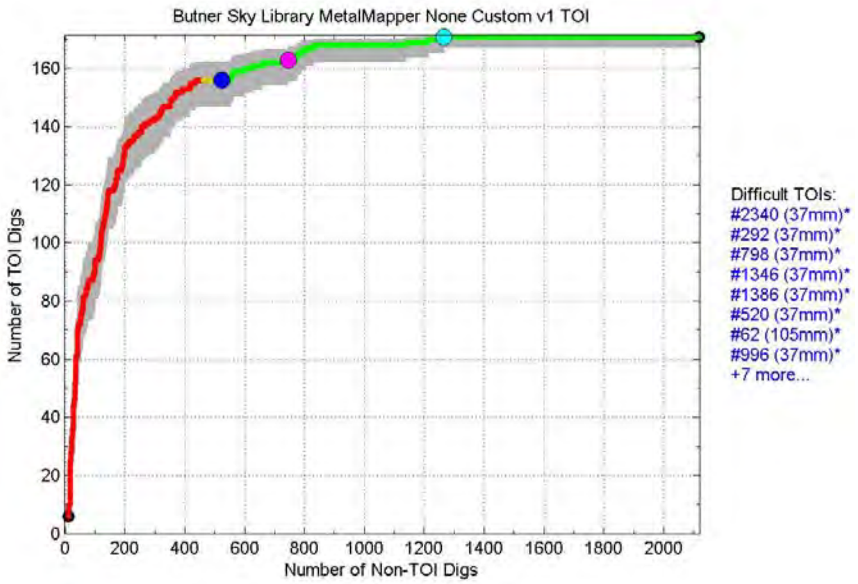


Figure 55: Sky's performance based on the sum and decay rates of polarizabilities inverted from the MetalMapper dynamic data using the UXOLab single-source and two-source dipole models and classified with a library matcher optimized over a custom training set.

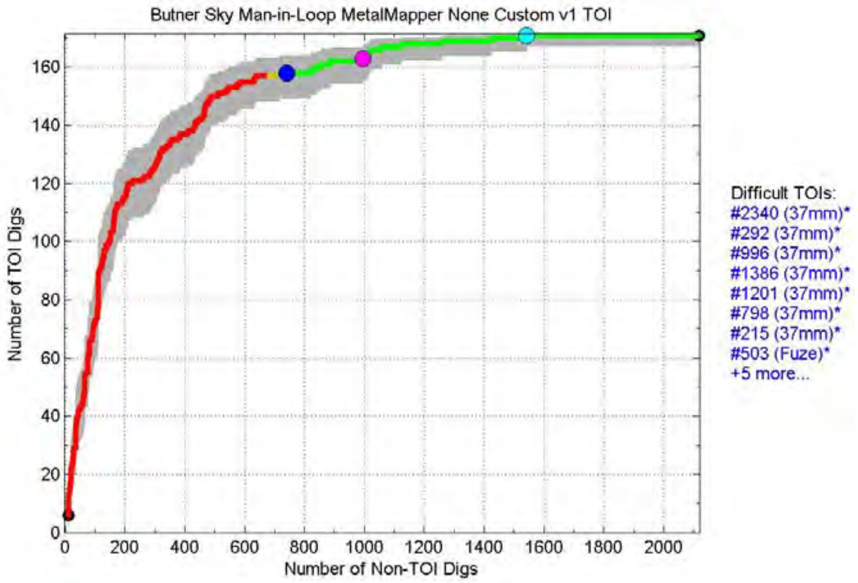


Figure 56: Sky's performance based on the sum and decay rates of polarizabilities inverted from the MetalMapper dynamic data using the UXOLab single-source and two-source dipole models and classified with a library matcher optimized over a custom training set, coupled with a human expert.

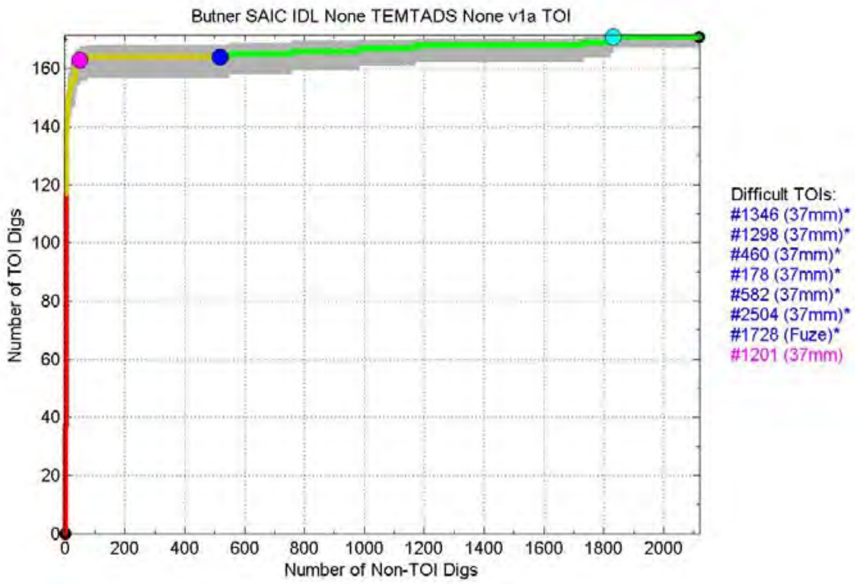


Figure 57: SAIC's performance based on amplitudes, ratios, and decay rates of polarizabilities inverted from the TEMTADS static data using an IDL single-source dipole model and classified with a rules-based library matcher optimized over an existing library.

A comparison of the two figures shows that the difference in dipole models did not affect the classification performance. The curves exhibit very similar shapes, indicating that the classification algorithms output very similar decision statistics and that the anomalies were ranked in very similar ways on the ranked anomaly lists. The one difference between the curves was the placement of the analyst's don't dig threshold. In the single-source analysis (Figure 57), seven TOIs were incorrectly classified at the analyst's don't dig threshold, while the number of Non-TOI digs was reduced by 1601 (76%, from 2120 to 519). In the multi-source analysis (Figure 48), three TOIs were incorrectly classified at this threshold, although the number of Non-TOI digs could be reduced by only 878 (41%, from 2120 to 1242). Different analysts performed these two analyses, however. The difference in the placement of the analyst's don't dig threshold is likely due to the more or less conservative nature of the analysts themselves, rather than any inherent differences in the estimated target characteristics.

2. Advanced Model

Dartmouth was the only team to use advanced, non-dipole geophysical models for inversion [28][56][59]. Targets that have a nonhomogeneous material composition or are in the near field of the sensor inspired these models. In these cases, the target heterogeneity and the sensor-target near-field effects are not well fit by a standard dipole model. In addition, the advanced, non-dipole models can be easily extended to handle overlapping anomalies caused by multiple, closely spaced targets. [34].

For each anomaly, the Dartmouth analyst estimated characteristics of the target using the advanced models and then input these characteristics into a statistical classifier trained over a small custom training set [55]. Results were excellent. Figure 58 and Figure 59 show the performance of the TEMTADS and MetalMapper static analyses, respectively. Both curves exhibit near-right angles, indicating that almost all Non-TOI anomalies were listed further up the ranked anomaly list than almost all TOI anomalies. In addition, the analysts' don't dig thresholds were nearly optimally chosen, placed very close to the retrospective "best" don't dig thresholds. In the TEMTADS analysis, the analyst's don't dig threshold correctly classified all TOIs while reducing the number of Non-TOI digs by 2004, a reduction of 95%. Similarly, the analyst's don't dig threshold also correctly classified all TOIs in the MetalMapper analysis, all while reducing the number of Non-TOI digs by 1946, a reduction of 92%. These were the two best results seen at the former Camp Butner.

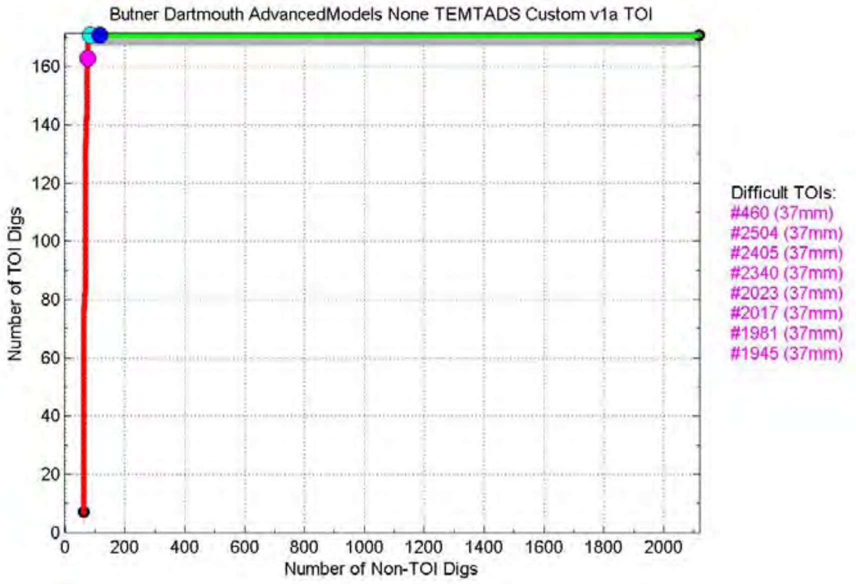


Figure 58: Dartmouth's performance based on full curves inverted from the TEMTADS static data using the OVNMS non-dipole model and classified with a statistical classifier optimized over a custom training set.

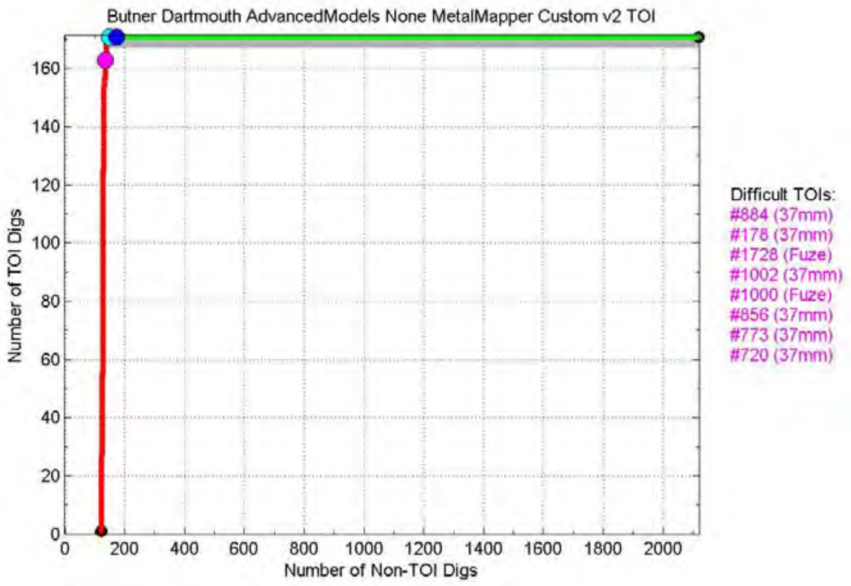


Figure 59: Dartmouth's performance based on full curves inverted from the MetalMapper static data using the OVNMS non-dipole model and classified with a statistical classifier optimized over a custom training set.

8. Conclusions

The results of the three successive live-site UXO classification demonstrations have built upon each other. The results from the first study, conducted at the former Camp Sibert, showed that good classification performance was possible using almost all types of data when conditions were benign, that is, when there was only one expected munition type and that munition was large compared with clutter. The results from the former Camp San Luis Obispo showed that classification was possible even when conditions were more challenging—that is, when multiple types of munitions were expected and some of those munitions were only slightly larger than clutter. In that case, high-quality data were necessary to achieve good classification performance. At the former Camp Butner, results showed that good classification performance was still possible in even more difficult conditions, those where many munitions were of the same size, or even smaller, than clutter. Both high-quality data and advanced processing techniques were needed to produce excellent results. The following sections summarize the findings of this demonstration and offer recommendations for future demonstrations.

A. Findings

1. **The EM61-Mk2 cart showed better detection performance than the MetalMapper in dynamic mode.** The EM61-Mk2 cart detected all seeded UXO, resulting in a Pd of 100% and a FAR of 487/acre. In comparison, the MetalMapper failed to detect two 37 mm projectiles seeded at 30 cm depths, resulting in a Pd of 99% and a FAR of 819/acre. Although the differences in Pd were not statistically significant, the MetalMapper's FAR was almost twice as high as the EM61-Mk2 cart's because its detection threshold was set much closer to the noise floor. Furthermore, had this been a real remediation project, stakeholders would have been troubled by the MetalMapper's inability to detect two seeds. The somewhat wider lane spacing used by the MetalMapper (0.75 m versus the 0.5 m for the EM61-Mk2) and the presence of one degraded MetalMapper triaxial receive coil may have contributed to the missed detections.

2. **The EM61-Mk2 cart exhibited poor classification performance.** The EM61-Mk2 analyses resulted in the incorrect classification of many TOIs and/or only a small reduction in the number of Non-TOI digs. This was likely because size is the only target characteristic that could be accurately estimated from the EM61-Mk2 data. Unfortunately, size was not a particularly useful feature at the former Camp Butner because so many TOIs were the same size as Non-TOIs.

3. **The MetalMapper in dynamic mode showed better classification performance than the EM61-Mk2 cart.** In comparison to the EM61-Mk2 analyses, the dynamic MetalMapper analyses often led to more TOIs correctly classified and/or greater reductions in the number of Non-TOI digs. The multiple, triaxial receive coils employed by the MetalMapper in dynamic mode likely led to a better relative cross-track position accuracy in the collected data, which led to more constrained estimates of the individual polarizabilities of the buried target. This, in turn, led to more accurate estimates of the target's size, as well as its shape and material composition/wall thickness. (The material composition/wall thickness estimate, however, was not a particularly discriminating feature for this instrument, given the short time window over which the received signal was sampled.)

4. **The MetalMapper in static mode provided more accurate classification than in dynamic mode.** Some static MetalMapper analyses led to the correct classification of most or all TOIs while reducing the number of Non-TOI digs by over 50%. This was likely due to three reasons:

   a. In static mode, the buried target was fully illuminated by all three orthogonal transmit coils, leading to more accurate estimates of the individual polarizabilities of the buried target, which in turn led to more accurate estimates of the target's size, shape, and material composition/wall thickness.

   b. The extended time gates used in static mode allowed a more accurate estimate of the decays of the target's polarizabilities at the later times where differences in material composition and wall thickness between TOIs and Non-TOIs become more evident.

   c. The static data used a larger stacking factor than the dynamic data, resulting in a higher SNR.

5. **The TEMTADS outperformed the MetalMapper in static mode.** Most TEMTADS analyses resulted in the correct classification of all TOIs while reducing the number of Non-TOI digs by more than 90%. This is likely due to the extended time gates used by the TEMTADS, as well as the careful field technique of the TEMTADS data-collection team.

6. **Classification algorithms did not have a large effect on classification performance, all other factors being equal.** Analyses based on the same data sets, geophysical inversions, feature selections, and training sets often led to very similar results, even when different classification algorithms were used (e.g., statistical classifier, library matcher, library matcher coupled with human expert, etc.).

7. **Partial static data requests were not helpful.** Analyses based on static data from only some anomalies exhibited much poorer performance than those based on static data from all anomalies, with what would be very little reduction in cost in a real remediation project. In some cases, TOIs were misclassified based solely on the EM61-Mk2 dynamic data, before the higher resolution static data were even requested.

8. **Custom training data requests were helpful.** In general, custom training sets were smaller than the Standard Training Set. This would reduce costs in a real remediation project, since fewer targets would have to be recovered to provide training data for algorithm optimization. Furthermore, custom training sets generally consisted of a larger percentage of TOIs than the Standard Training Set because the classification analyst teams often requested ground truth on representative anomalies from each cluster in multidimensional feature space.

9. **Commercial geophysics companies performed well.** The commercial companies were mentored by organizations with more experience in UXO classification. In two cases, the commercial companies outperformed their mentors. This may have been because the commercial companies focused their time and resources on fewer types of analyses.

10. **Second-pass analyses led to improvements in classification performance.** Two classification analyst teams refined their classification algorithms once ground truth became known for anomalies that were dug based on the first pass of analyses. This more closely mimics what could occur in a real remediation project. Second-pass analyses, coupled with UXO seeding, could allow stakeholders to have more confidence in the final don’t dig threshold.

11. **The number of dipole sources assumed by geophysical inversion routines did not have a large effect on classification performance.** Analyses based on the same data sets, dipole models, feature selections, classification algorithms, and training sets often led to very similar results, even when different assumptions were made about the number of dipole sources per anomaly.

12. **Advanced geophysical models led to excellent results.** These models were inspired by nonhomogeneous targets in the near field of the sensor, as well as multiple, closely spaced targets leading to overlapping anomalies. These models assumed multiple non-dipole sources for each anomaly. The former Camp Butner was a challenging enough site such that high-quality data alone did not consistently lead to excellent results—advanced data-processing techniques were also necessary.

B. Recommendations

1. **Sufficient time and resources for quality control should be built in to future demonstration plans.** At the former Camp Butner, quality control checks at ESTCP and IDA promptly caught problems related to data collection and anomaly detection, such that those problems could be swiftly addressed. This is especially important for dual-mode instruments like the MetalMapper, in which stakeholders may feel tempted to schedule little time between collecting dynamic and static data.

2. **Custom training data requests should be used in future demonstrations.** At the former Camp Butner, these requests led to smaller and more TOI-laden training sets than the Standard Training Set, which would reduce costs in a real remediation project. The custom training data requests also allowed the classification analysts teams to identify the ground-truth labels of unknown clusters in multidimensional feature space.

3. **The classification analysis teams should be given the opportunity to perform multiple-pass analyses in future demonstrations.** In the first few passes, ESTCP could provide feedback to the analyst regarding which seeded UXO were incorrectly classified at the analyst's most recent don't dig threshold. In essence, these seeded UXO would be added to the training set for the next pass of analysis. Once all seeded UXO were correctly classified, ESTCP could then begin providing feedback regarding the ground-truth labels of all anomalies that fell below the analyst's most recent don't dig threshold. The analyst could use this ground truth to re-optimize his or her classification algorithms and then apply these re-optimized algorithms to only those anomalies that rose above the most recent don't dig threshold. This would mimic what could occur in a real remediation project, where feedback over multiple passes could give stakeholders more confidence in the final don't dig threshold.

4. **Future demonstrations should seed UXO in randomly selected locations.** In the past three demonstrations, IDA took care to select seed locations that were far from large anomalies representing native targets because classification technologies had not yet been shown to properly address overlapping anomalies. Advanced geophysical models are now being developed to address this issue.

5. **Future demonstrations should rank anomalies in reverse order on the ranked anomaly lists.** The first anomaly on the list should be the first anomaly that is dug (i.e., the anomaly most likely to be a TOI). This would reduce confusion during tech transfer to stakeholders in real remediation projects.

6. **Classification analyst teams should be limited to only a handful of different types of classification analyses in future demonstrations.** At the former Camp

Butner, commercial geophysics companies with little to no experience in UXO classification outperformed their mentors, even though their mentors had several years of experience in UXO classification. This is likely because the mentors performed too many types of analyses, spreading their time and resources too thin.

7. **More commercial geophysics companies should be encouraged to take part in future demonstrations.** Participation in the demonstrations will be an excellent opportunity for training in UXO classification technology. This will jump-start tech transfer.

Appendix A. Classification Performance Curves

EM61-Mk2 Dynamic Data

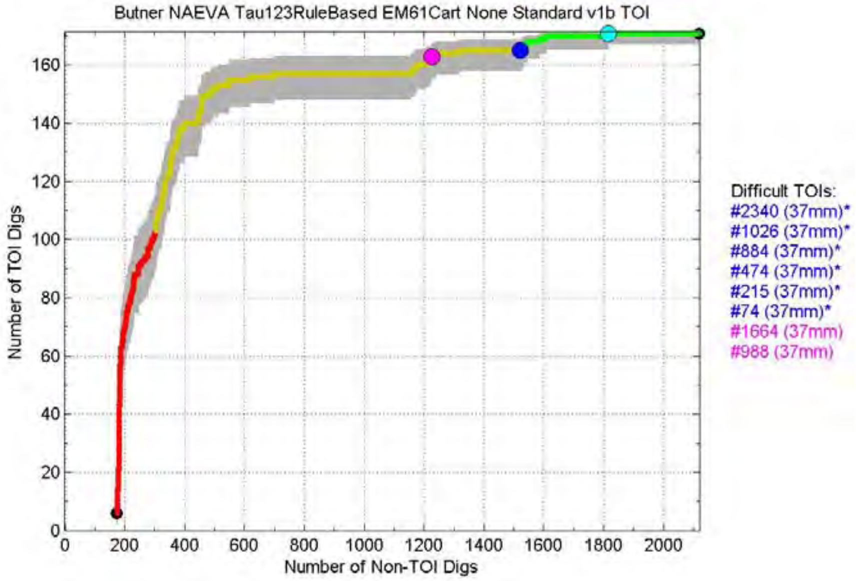


Figure 60: NAEVA's performance based on the decay rates of the peak amplitudes of the anomalies, measured from the EM61-Mk2 dynamic data with UX-Detect and UX-Process and classified with rules optimized over the standard training set. Identical to Figure 39 in the text.

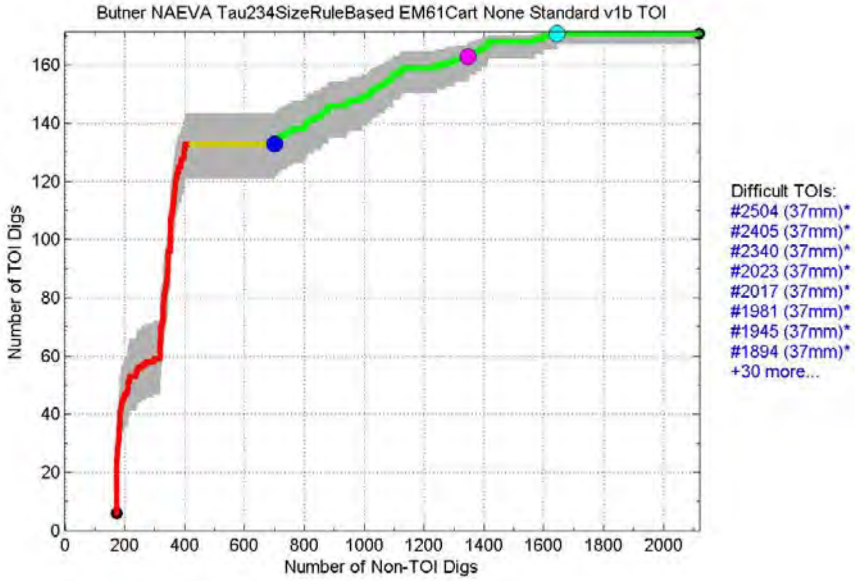


Figure 61: NAEVA's performance based on the decay rates and footprints of the anomalies, measured from the EM61-Mk2 dynamic data with UX-Detect and UX-Process and classified with rules optimized over the standard training set.

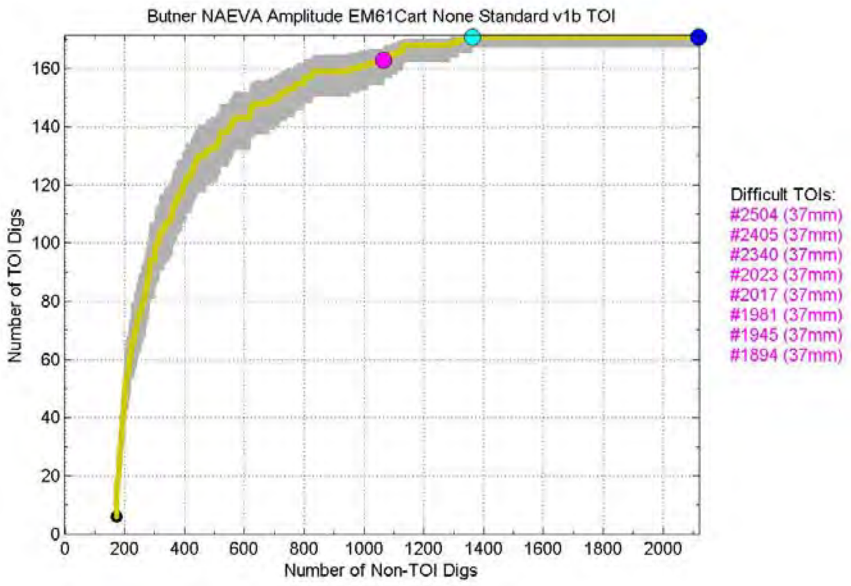


Figure 62: NAEVA's performance based on the peak amplitudes of the anomalies, measured from the EM61-Mk2 dynamic data with UX-Detect and UX-Process and classified with rules optimized over the standard training set. All anomalies in the standard test set were declared "Can't Decide" (yellow) because the don't dig threshold could not be set prospectively. Identical to Figure 40 in the text.

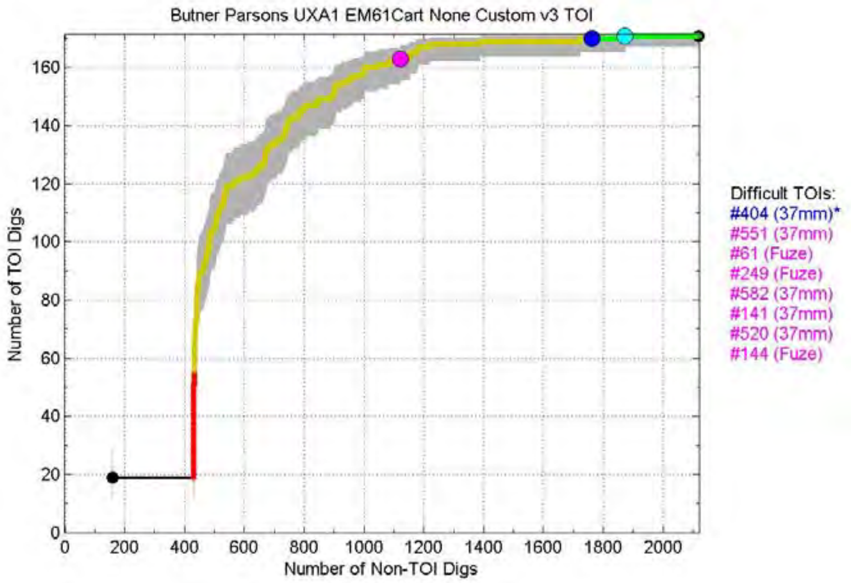


Figure 63: Parson's performance for its first retrospective analysis based on the decay rates of polarizabilities inverted from the EM61-Mk2 dynamic data using the UX-Analyze dipole model and classified with rules optimized over a custom training set.

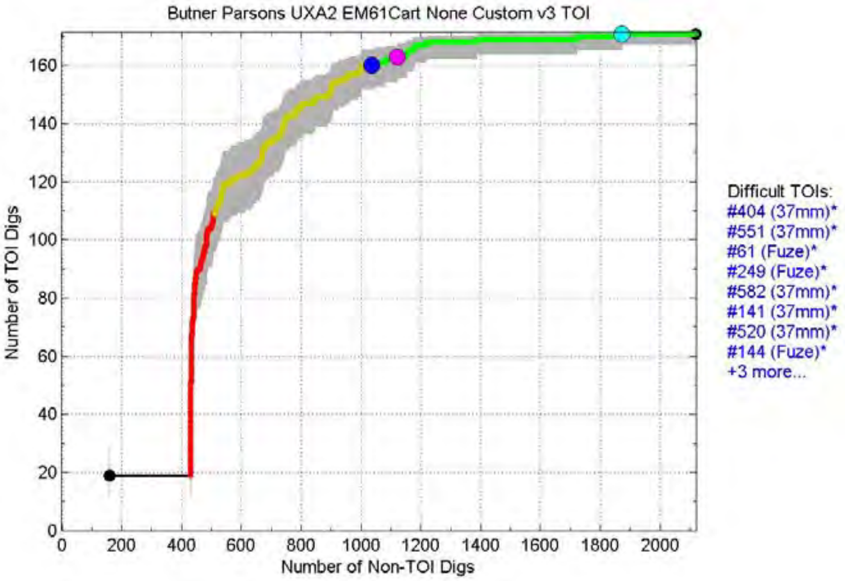


Figure 64: Parson's performance for its second retrospective analysis based on the decay rates of polarizabilities inverted from the EM61-Mk2 dynamic data using the UX-Analyze dipole model and classified with rules optimized over a custom training set.

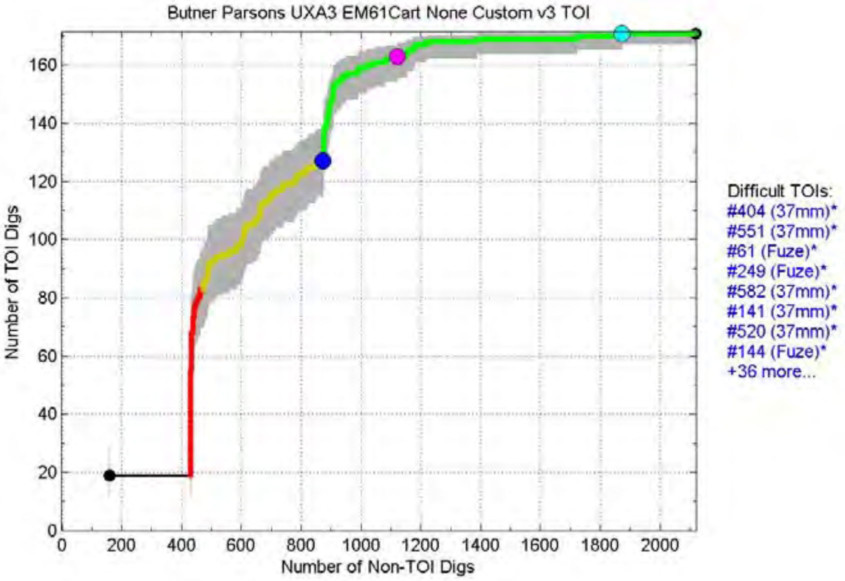


Figure 65: Parson's performance for its third retrospective analysis based on the decay rates of polarizabilities inverted from the EM61-Mk2 dynamic data using the UX-Analyze dipole model and classified with rules optimized over a custom training set.

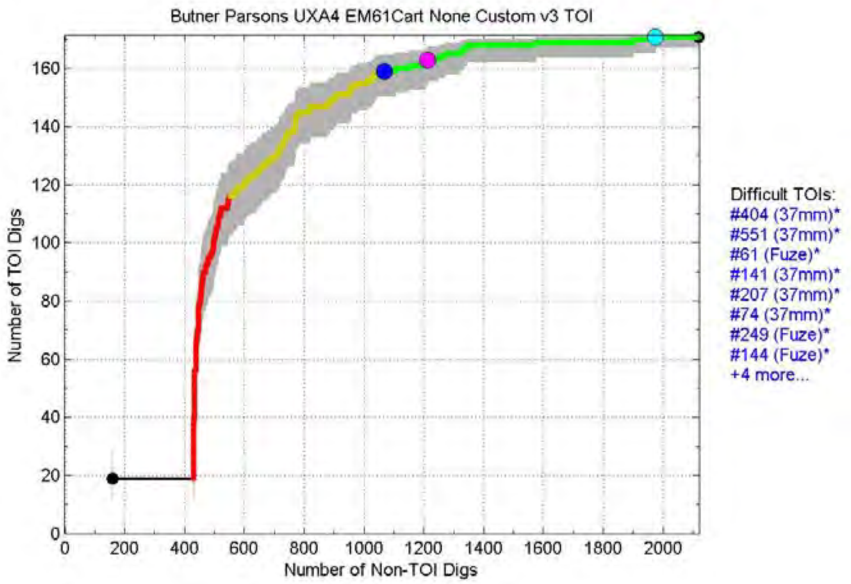


Figure 66: Parson's performance for its fourth retrospective analysis based on the decay rates of polarizabilities inverted from the EM61-Mk2 dynamic data using the UX-Analyze dipole model and classified with rules optimized over a custom training set.

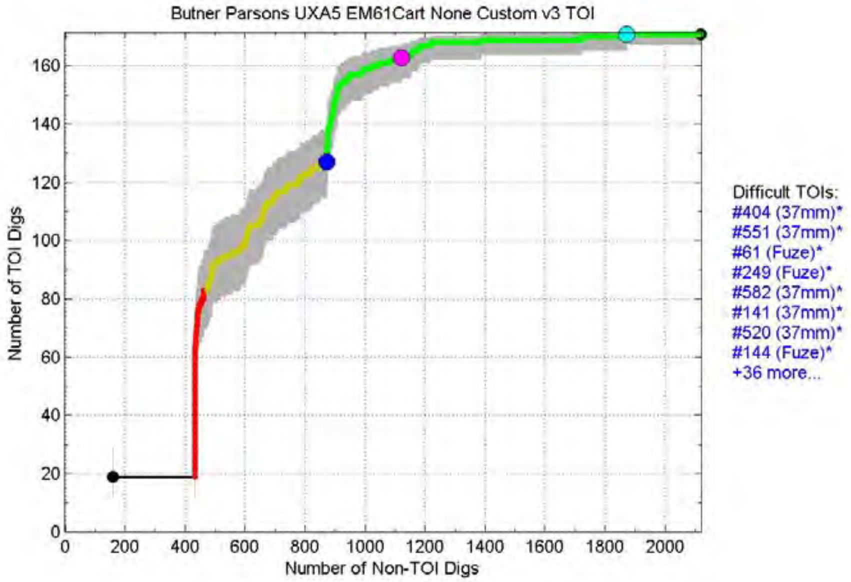


Figure 67: Parson's performance for its fifth retrospective analysis based on the decay rates of polarizabilities inverted from the EM61-Mk2 dynamic data using the UX-Analyze dipole model and classified with rules optimized over a custom training set.

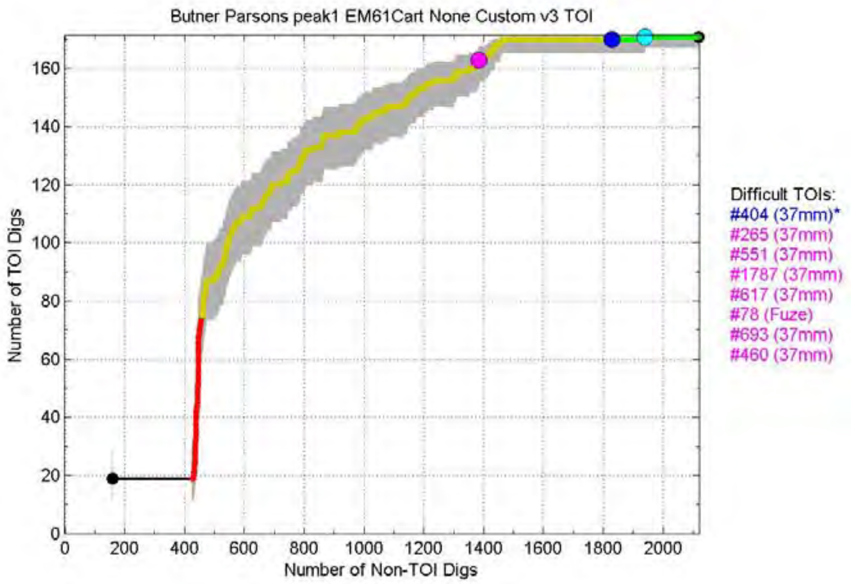


Figure 68: Parson's performance for its sixth retrospective analysis based on the decay rates of polarizabilities inverted from the EM61-Mk2 dynamic data using the UX-Analyze dipole model and classified with rules optimized over a custom training set.

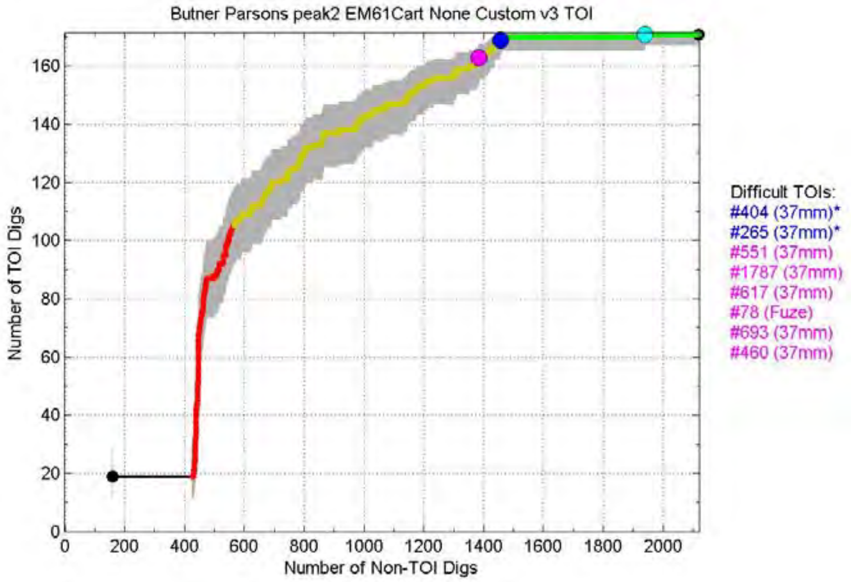


Figure 69: Parson's performance for its seventh retrospective analysis based on the decay rates of polarizabilities inverted from the EM61-Mk2 dynamic data using the UX-Analyze dipole model and classified with rules optimized over a custom training set.

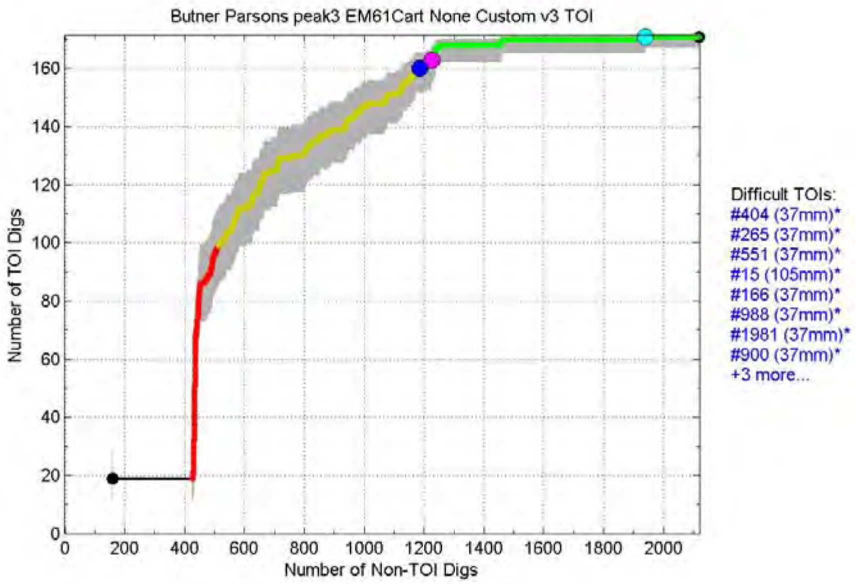


Figure 70: Parson's performance for its eighth retrospective analysis based on the decay rates of polarizabilities inverted from the EM61-Mk2 dynamic data using the UX-Analyze dipole model and classified with rules optimized over a custom training set.

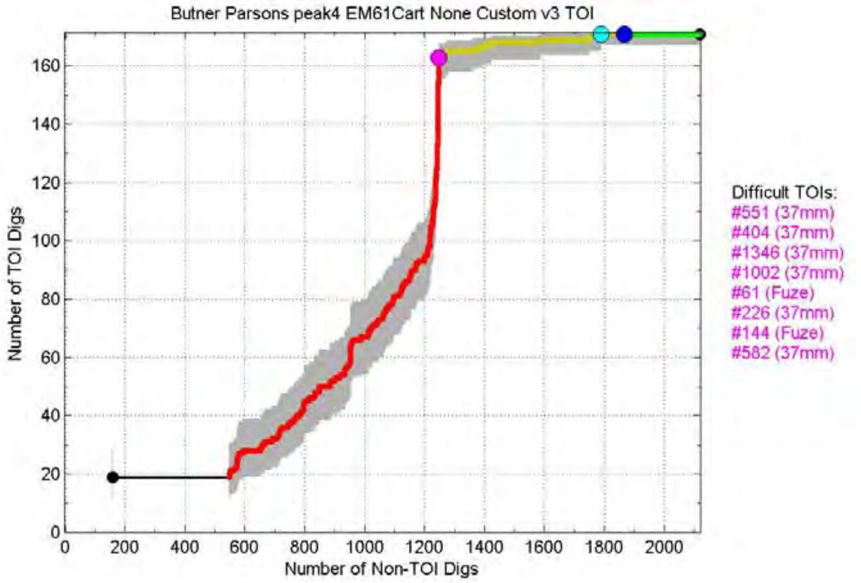


Figure 71: Parson's performance for its ninth retrospective analysis based on the decay rates of polarizabilities inverted from the EM61-Mk2 dynamic data using the UX-Analyze dipole model and classified with rules optimized over a custom training set.

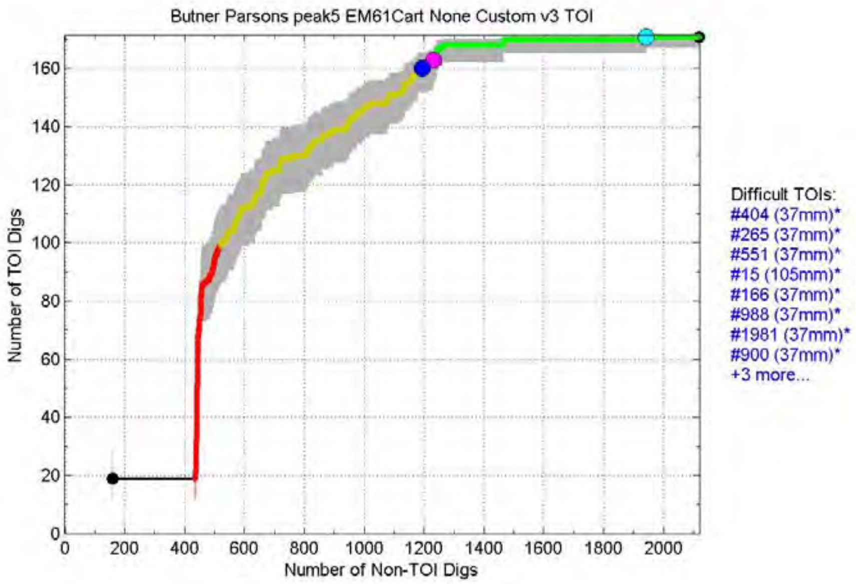


Figure 72: Parson's performance for its 10th retrospective analysis based on the decay rates of polarizabilities inverted from the EM61-Mk2 dynamic data using the UX-Analyze dipole model and classified with rules optimized over a custom training set.

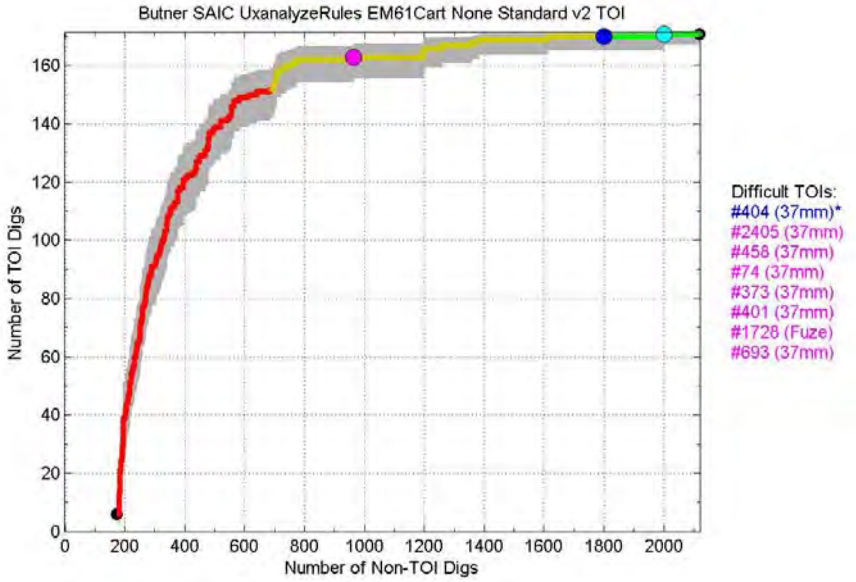


Figure 73: SAIC's performance for a retrospective analysis based on amplitudes and decay rates of polarizabilities inverted from the EM61-Mk2 dynamic data using the UX-Analyze/IDL multi-source dipole model and classified with a rules-based library matcher optimized over the standard training set. Identical to Figure 41 in the text.

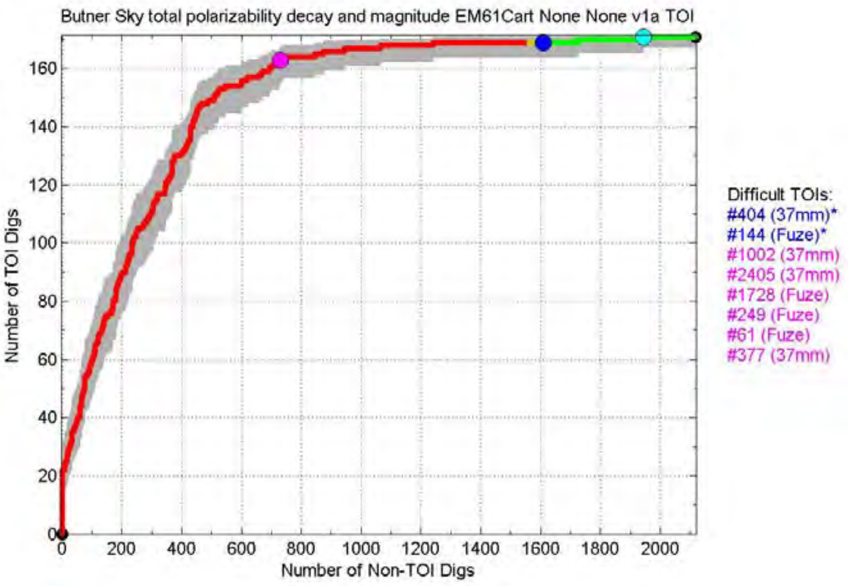


Figure 74: Sky's performance based on amplitudes and decay rates of polarizabilities inverted from the EM61-Mk2 dynamic data using the UXOLab single-source and two-source dipole models and classified with a statistical classifier optimized over an existing library. Identical to Figure 42 in the text.

MetalMapper Dynamic Data

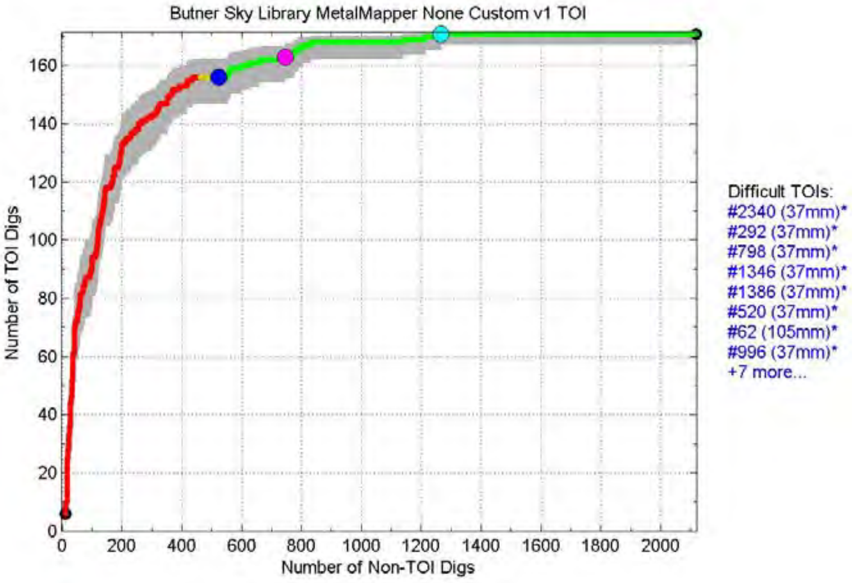


Figure 75: Sky's performance based on amplitudes and decay rates of polarizabilities inverted from the MetalMapper dynamic data using the UXOLab single-source and two-source dipole models and classified with a library matcher optimized over a custom training set. Identical to Figure 55 in the text.

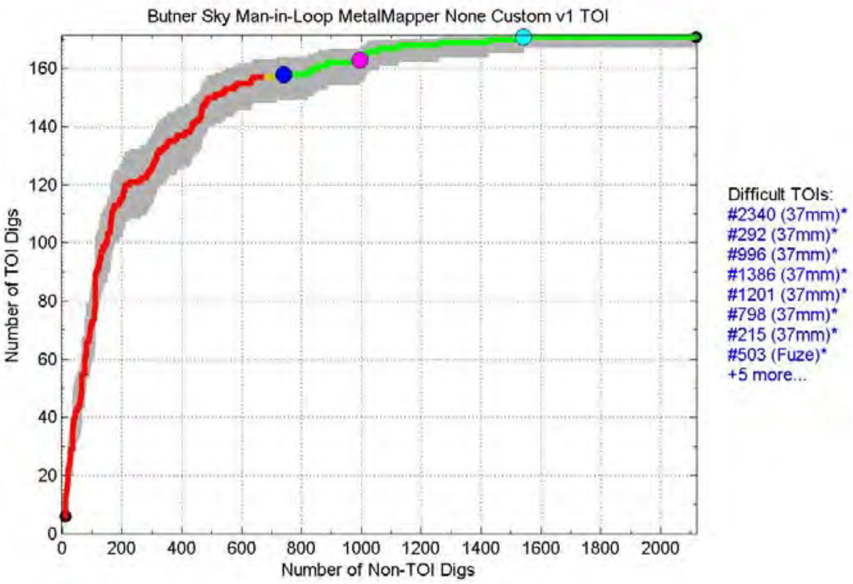


Figure 76: Sky's performance based on the sum and decay rates of polarizabilities inverted from the MetalMapper dynamic data using the UXOLab single-source and two-source dipole models and classified with a library matcher optimized over a custom training set, coupled with a human expert. Identical to Figure 56 in the text.

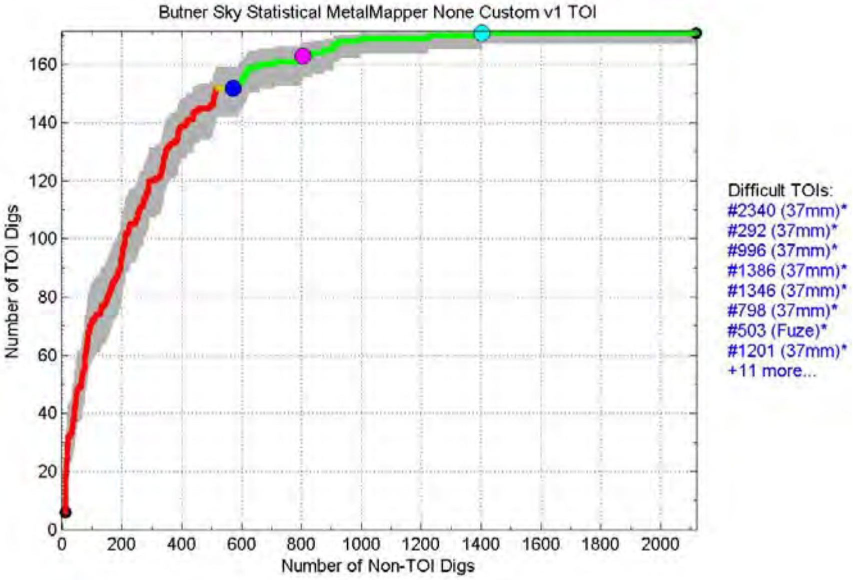


Figure 77: Sky's performance based on amplitudes and decay rates of polarizabilities inverted from the MetalMapper dynamic data using the UXOLab single-source and two-source dipole models and classified with a statistical classifier optimized over a custom training set. Identical to Figure 43 in the text.

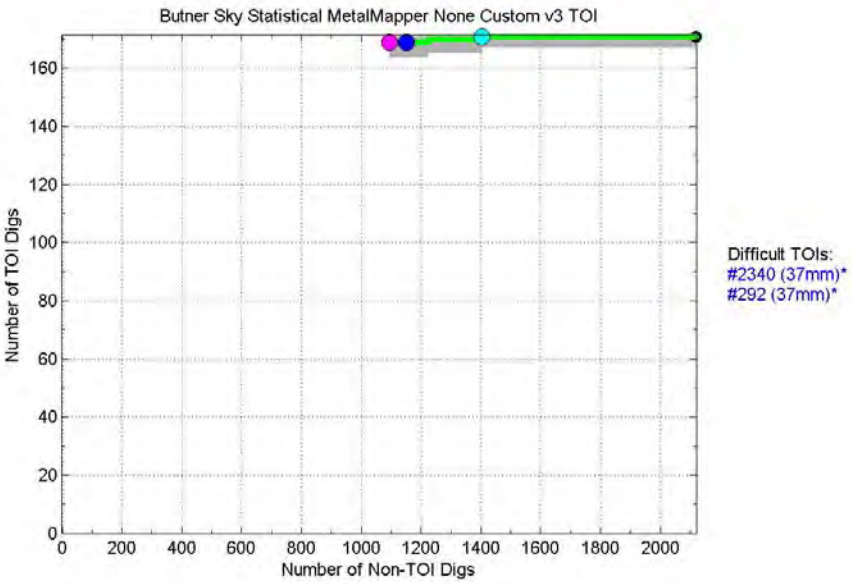


Figure 78: Sky's performance based on amplitudes and decay rates of polarizabilities inverted from the MetalMapper dynamic data using the UXOLab single-source and two-source dipole models and classified with a statistical classifier optimized over a custom training set. This second-pass custom training set consisted of those anomalies that fell below the don't dig threshold in the previous figure, as well as (in error) some anomalies that rose above threshold.

MetalMapper Static Data

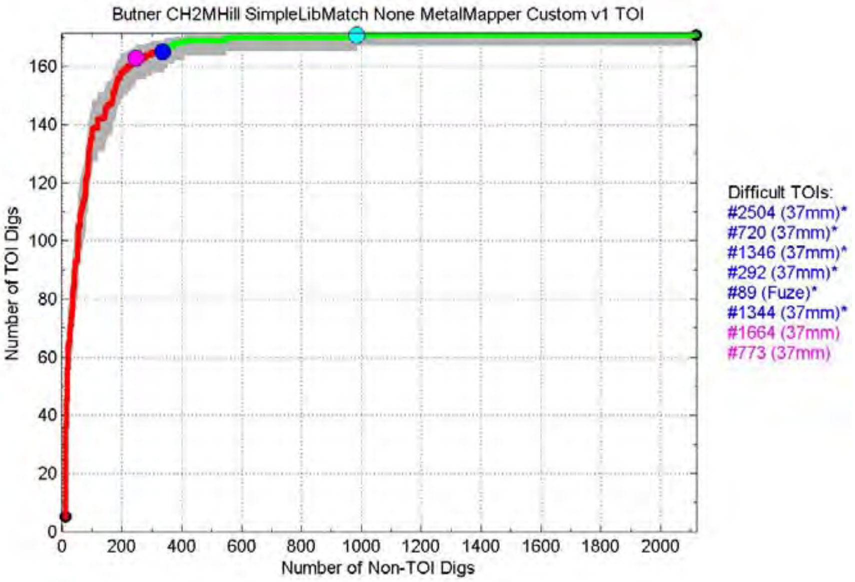


Figure 79: CH2M HILL's performance based on polarizabilities inverted from the MetalMapper static data using the UXOLab single-source and two-source dipole models and classified with a library matcher optimized over a custom training set. Identical to Figure 54 in the text.

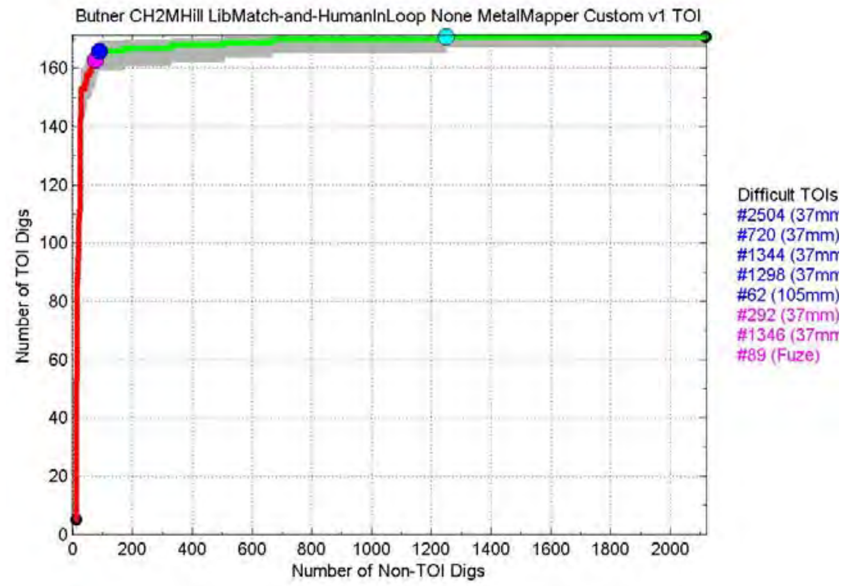


Figure 80: CH2M HILL's performance based on polarizabilities inverted from the MetalMapper static data using the UXOLab single-source and two-source dipole models and classified with a library matcher optimized over a custom training set, coupled with a human expert.

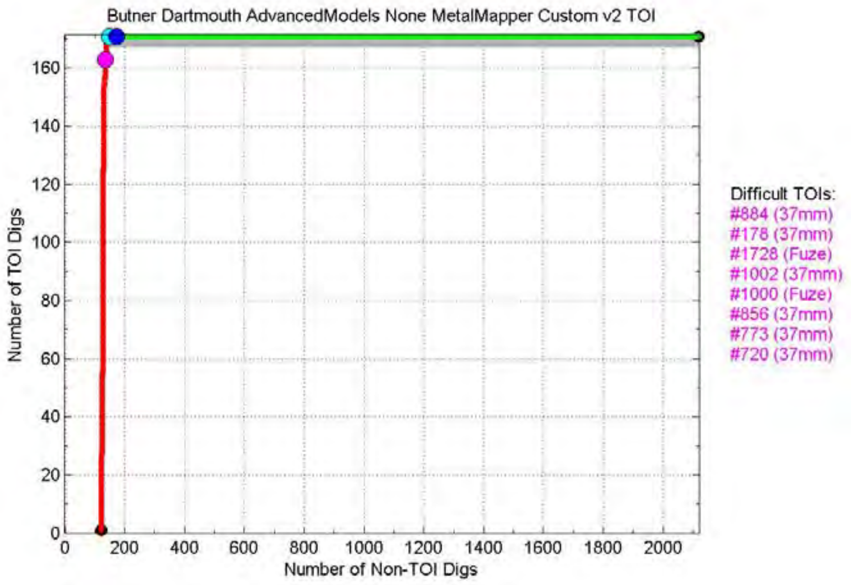


Figure 81: Dartmouth's performance based on full curves inverted from the MetalMapper static data using the OVNMS non-dipole model and classified with a statistical classifier optimized over a custom training set. Identical to Figure 59 in the text.

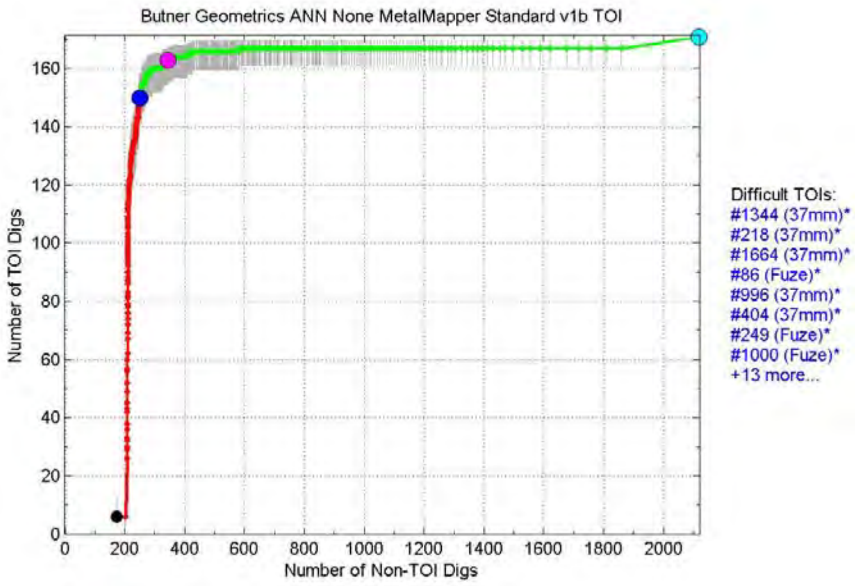


Figure 82: Geometrics' performance based on polarizabilities inverted from the MetalMapper static data using the MMRMP single-source dipole model and classified with an artificial neural network optimized over the standard training set.

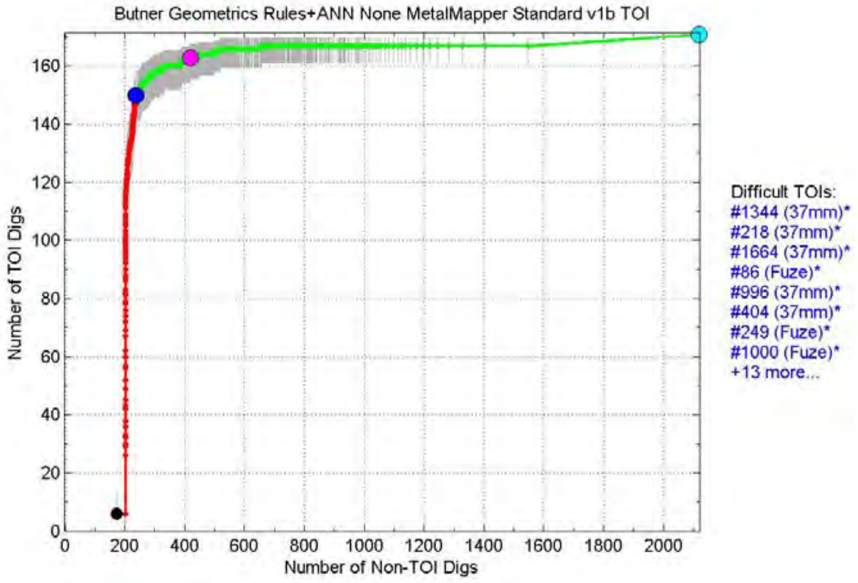


Figure 83: Geometrics' performance based on polarizabilities inverted from the MetalMapper static data using the MMRMP single-source dipole model and classified with an artificial neural network and rules optimized over the standard training set.

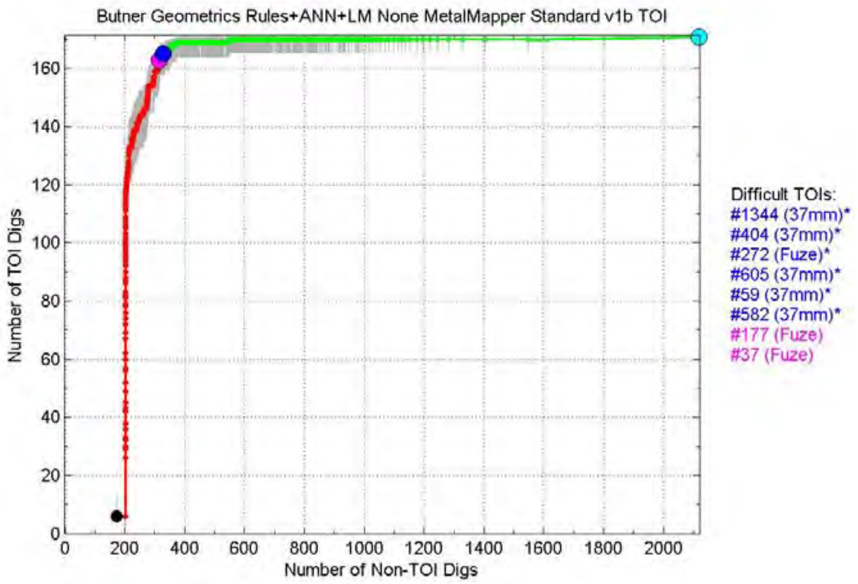


Figure 84: Geometrics' performance based on polarizabilities inverted from the MetalMapper static data using the MMRMP single-source dipole model and classified with an artificial neural network, rules, and a library matcher optimized over the standard training set.

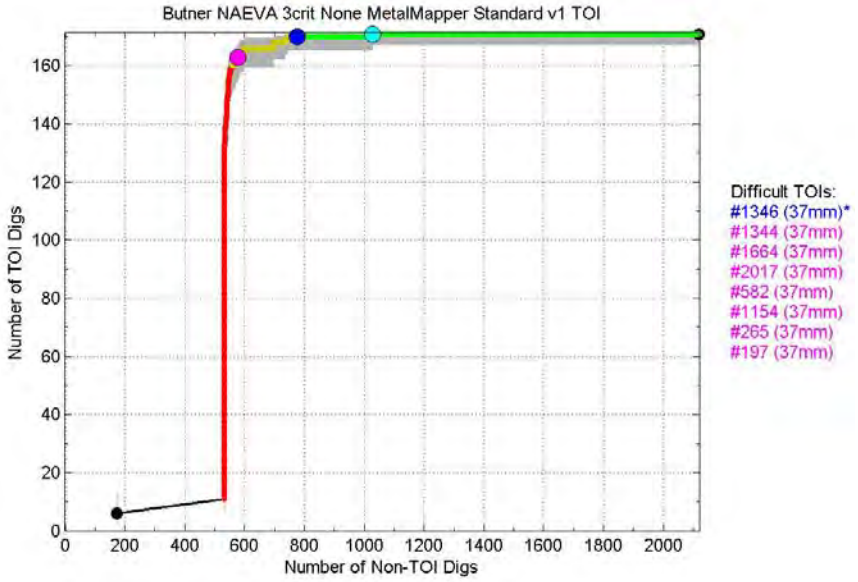


Figure 85: NAEVA's performance for a retrospective analysis based on polarizabilities inverted from the MetalMapper static data using the UX-Analyze single-source and multi-source dipole model and classified with a rules-based library matcher optimized over the standard training set. Only one stage of rules was used.

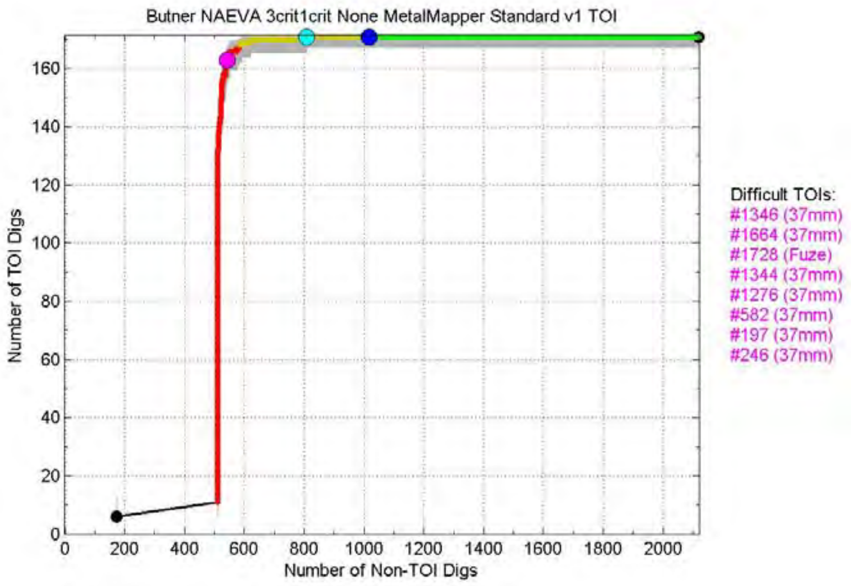


Figure 86: NAEVA's performance for a retrospective analysis based on polarizabilities inverted from the MetalMapper static data using the UX-Analyze single-source and multi-source dipole model and classified with a rules-based library matcher optimized over the standard training set. Two stages of rules were used; the second stage used only the polarizability amplitudes.

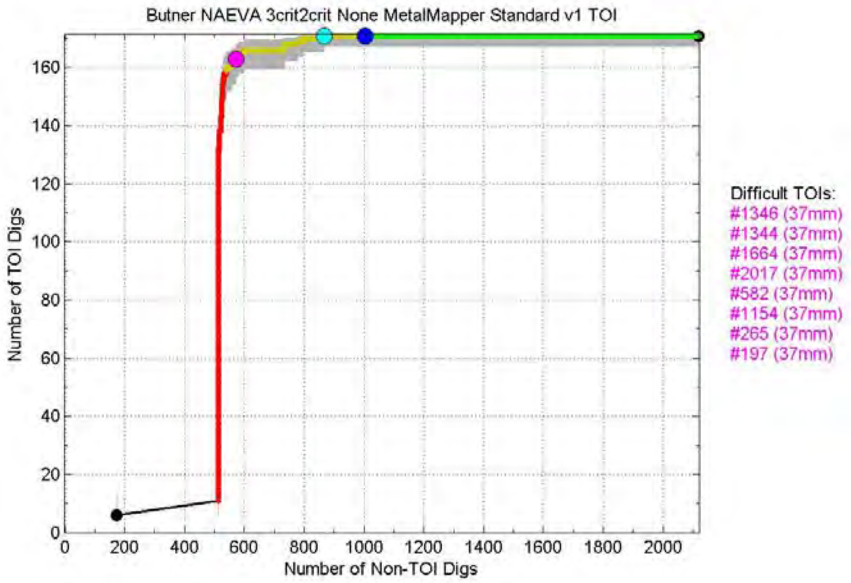


Figure 87: NAEVA's performance for a retrospective analysis based on polarizabilities inverted from the MetalMapper static data using the UX-Analyze single-source and multi-source dipole model and classified with a rules-based library matcher optimized over the standard training set. Two stages of rules were used; the second stage used only the polarizability amplitudes and ratios. Identical to Figure 52 in the text.

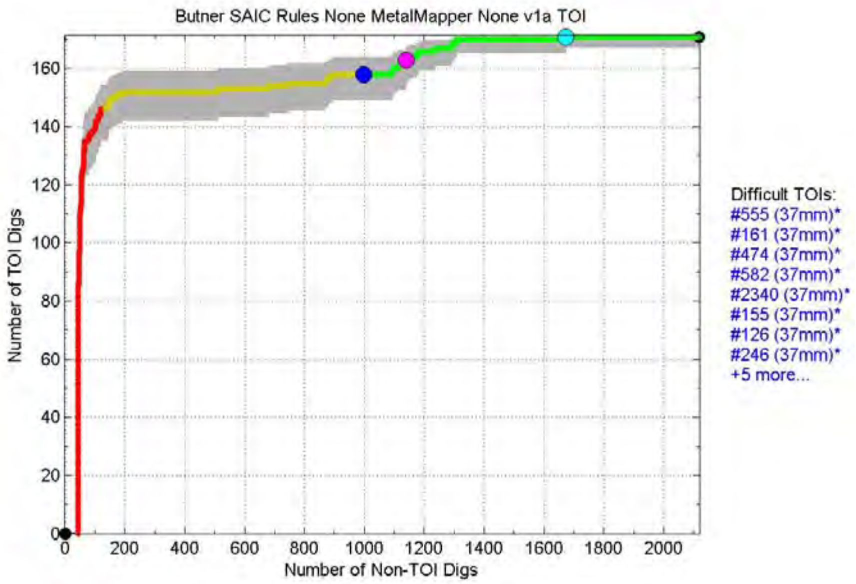


Figure 88: SAIC's performance based on amplitudes, ratios, and decay rates of polarizabilities inverted from the MetalMapper static data using an IDL single-source dipole model and classified with a rules-based library matcher optimized over an existing library. Identical to Figure 50 in the text.

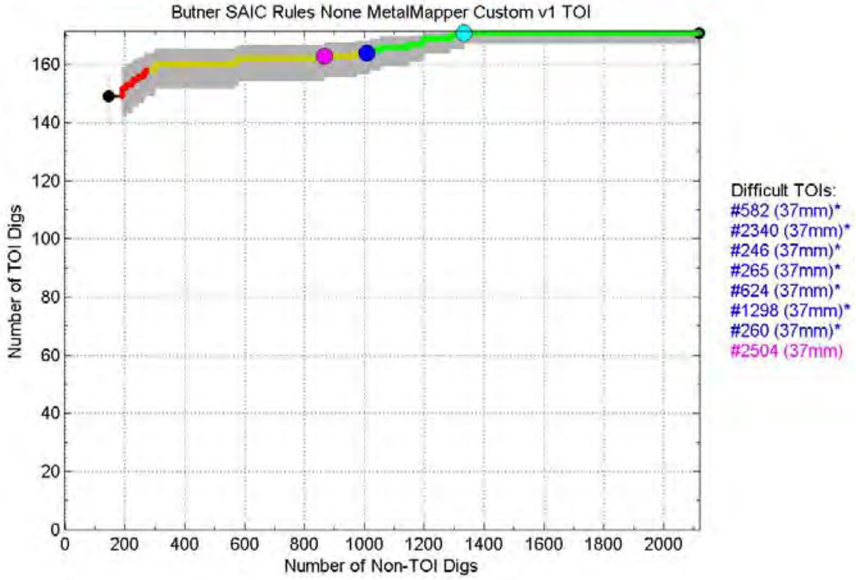


Figure 89: SAIC's performance based on amplitudes, ratios, and decay rates of polarizabilities inverted from the MetalMapper static data using an IDL single-source dipole model and classified with a rules-based library matcher optimized over a custom training set. This second-pass custom training set was built from the "Likely TOI" anomalies in the previous figure. Identical to Figure 51 in the text.

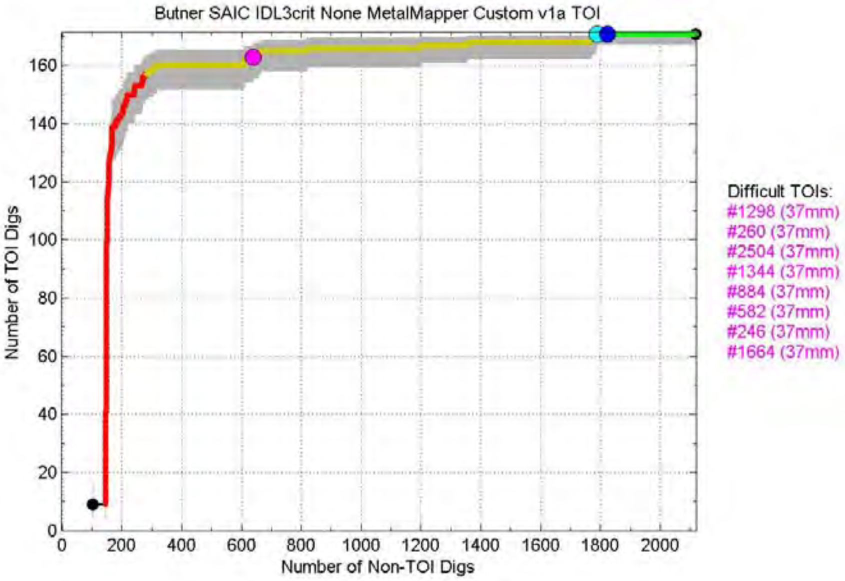


Figure 90: SAIC's performance based on amplitudes, ratios, and decay rates of polarizabilities inverted from the MetalMapper static data using an IDL single-source dipole model and classified with a rules-based library matcher optimized over a custom training set. Two polarizability ratios were used.

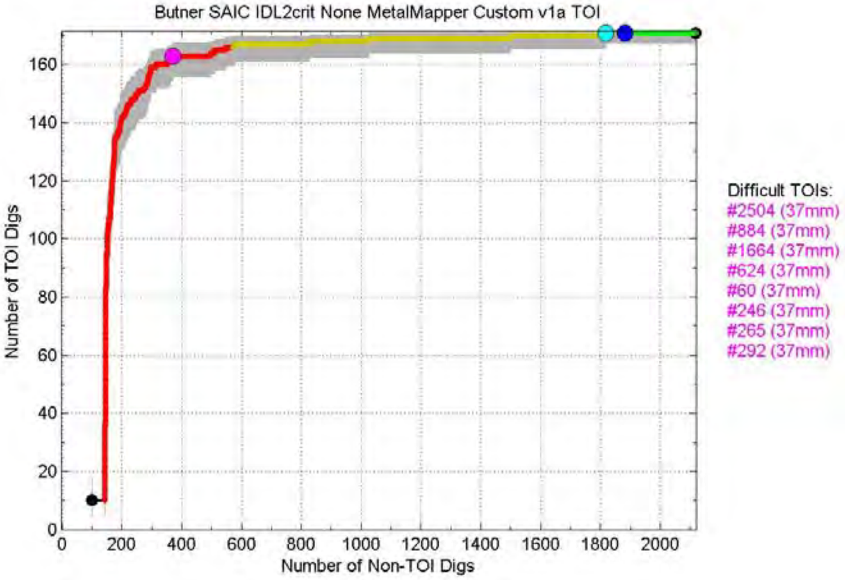


Figure 91: SAIC's performance based on amplitudes, ratios, and decay rates of polarizabilities inverted from the MetalMapper static data using an IDL single-source dipole model and classified with a rules-based library matcher optimized over a custom training set. One polarizability ratio was used.

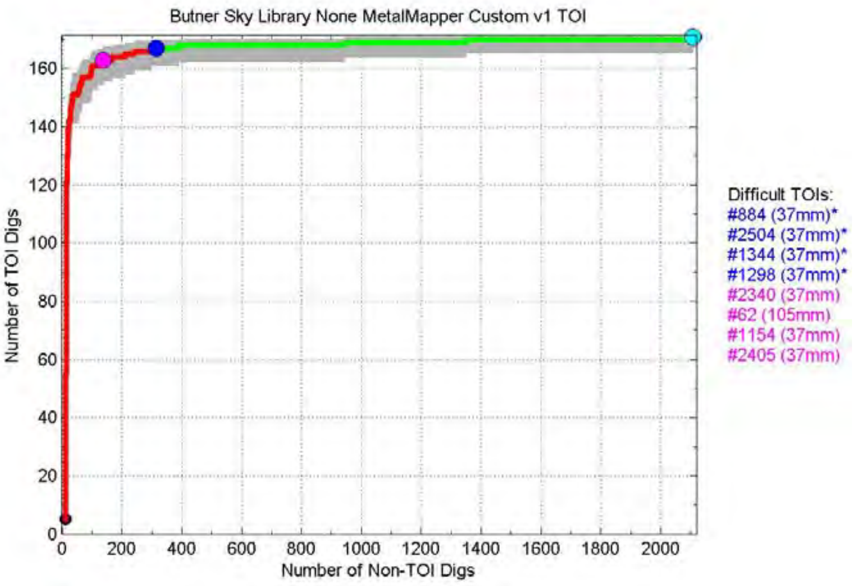


Figure 92: Sky's performance based on full polarizability curves inverted from the MetalMapper static data using the UXOLab single-source and two-source dipole models and classified with a library matcher optimized over a custom training set. Identical to Figure 53 in the text.

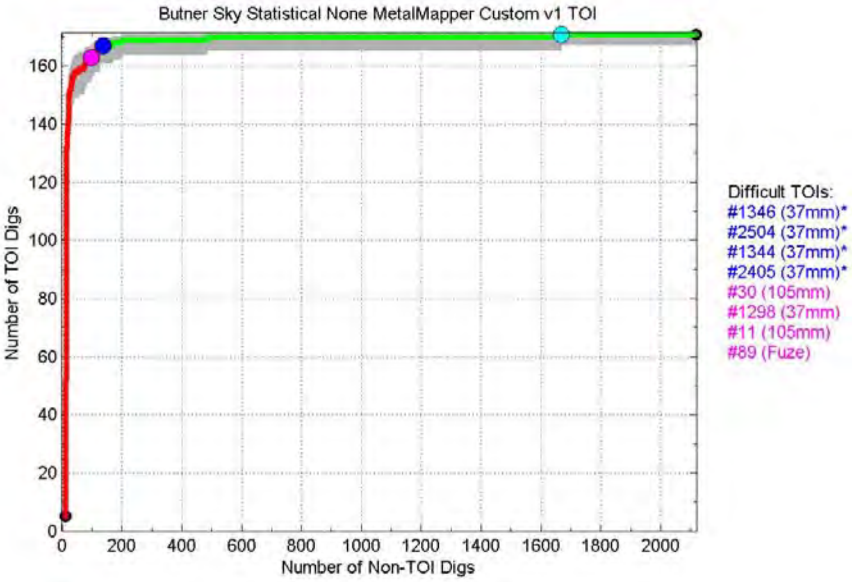


Figure 93: Sky's performance based on full polarizability curves inverted from the MetalMapper static data using the UXOLab single-source and two-source dipole models and classified with a statistical classifier optimized over a custom training set. Identical to Figure 44 in the text.

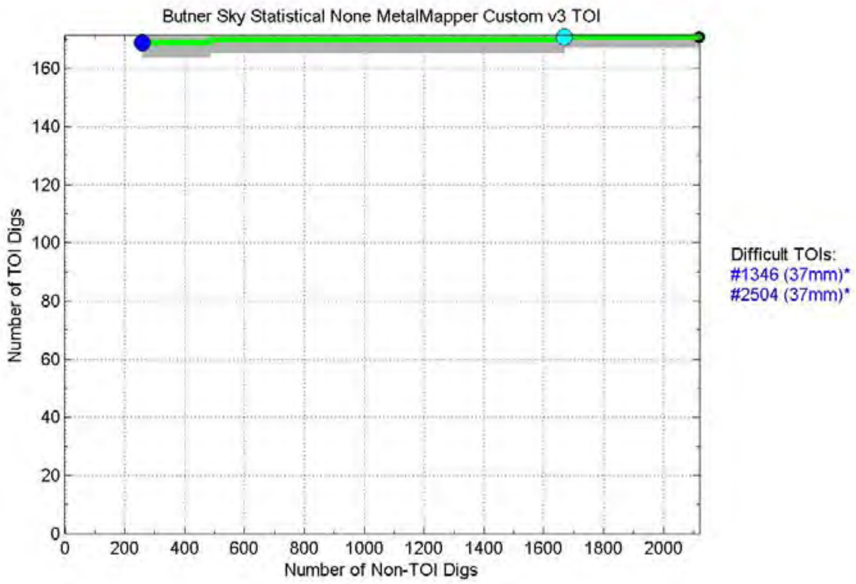


Figure 94: Sky's performance based on full polarizability curves inverted from the MetalMapper static data using the UXOLab single-source and two-source dipole models and classified with a statistical classifier optimized over a custom training set. This second-pass custom training set consisted of those anomalies that fell below the don't dig threshold in the previous figure, as well as (in error) some anomalies that rose above threshold.

EM61-Mk2 Dynamic Data with MetalMapper Static Data Requests

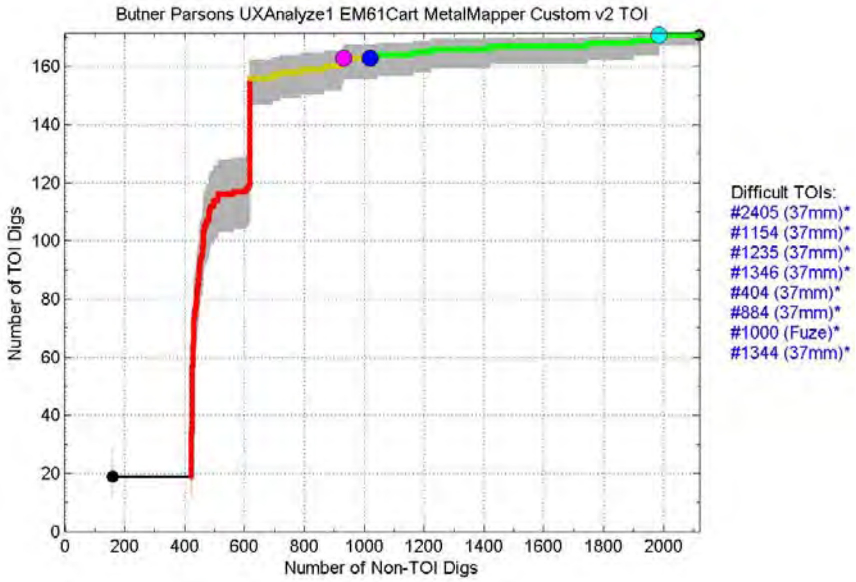


Figure 95: Parsons' performance based on polarizabilities inverted from the EM61-Mk2 dynamic data and requested MetalMapper static data using the UX-Analyze multisource dipole model and classified with a library matcher optimized over an existing library. The library included all targets (TOI and Non-TOI) built in to UX-Analyze.

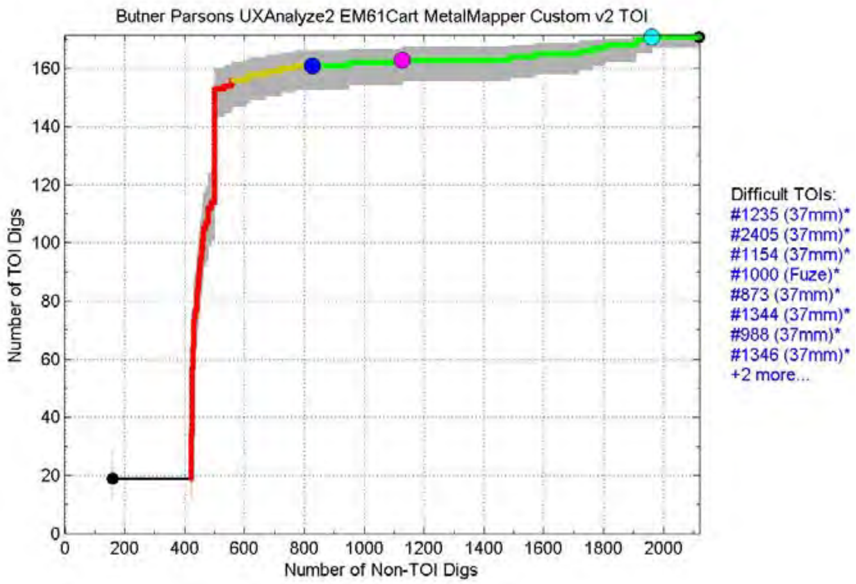


Figure 96: Parsons' performance based on polarizabilities inverted from the EM61-Mk2 dynamic data and requested MetalMapper static data using the UX-Analyze multisource dipole model and classified with a library matcher optimized over an existing library. The library included all targets (TOI and Non-TOI) built in to UX-Analyze, except for 20 mm projectiles.

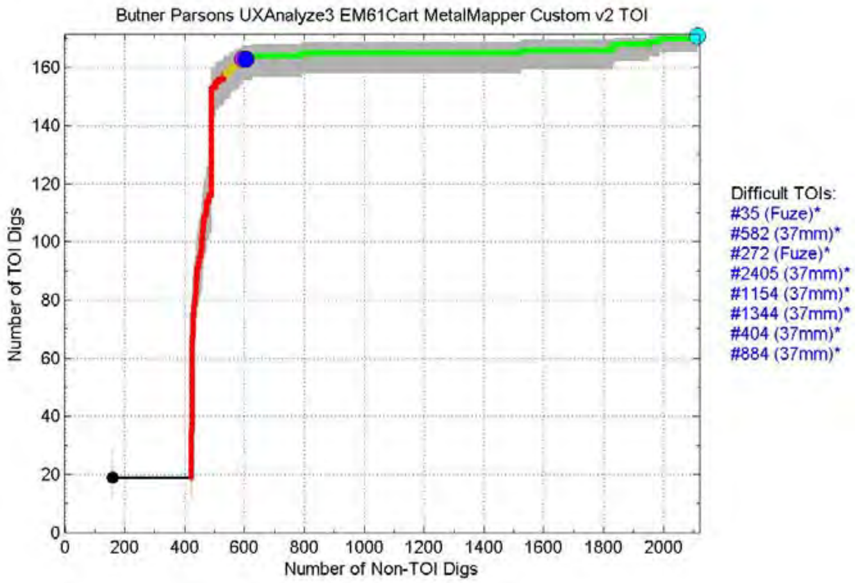


Figure 97: Parsons' performance based on polarizabilities inverted from the EM61-Mk2 dynamic data and requested MetalMapper static data using the UX-Analyze multisource dipole model and classified with a rules-based library matcher optimized over an existing library. The library included M48 fuzes, 2.36 in rockets, 37 mm projectiles, 40 mm projectiles, 81 mm mortars, 105 mm projectiles, and 155 mm projectiles, as well as all Non-TOIs built in to UX-Analyze.

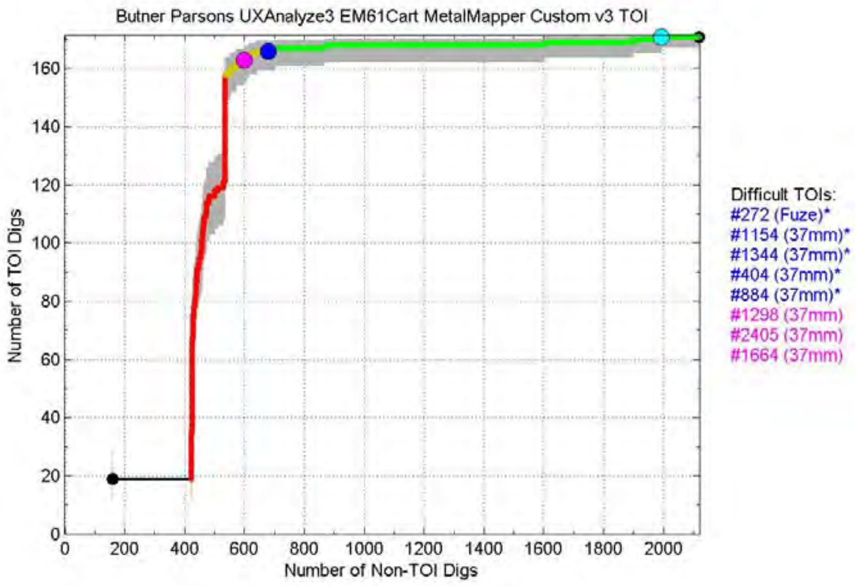


Figure 98: Parsons' performance based on a revised version of the analysis illustrated in the previous figure.

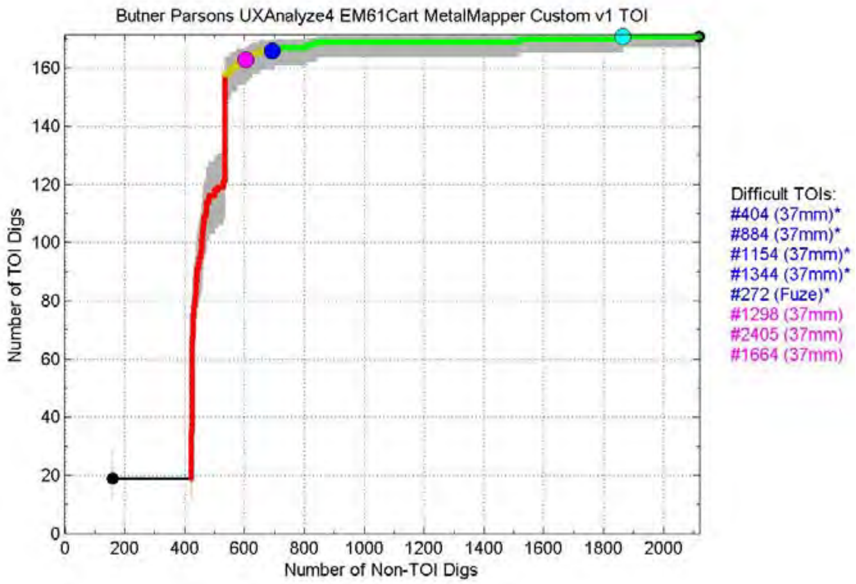


Figure 99: Parsons' performance based on polarizabilities inverted from the EM61-Mk2 dynamic data and requested MetalMapper static data using the UX-Analyze multisource dipole model and classified with a library matcher optimized over an existing library. The library included M48 fuzes, 2.36 in rockets, 37 mm projectiles, 40 mm projectiles, 81 mm mortars, 105 mm projectiles, and 155 mm projectiles; no Non-TOIs were included.

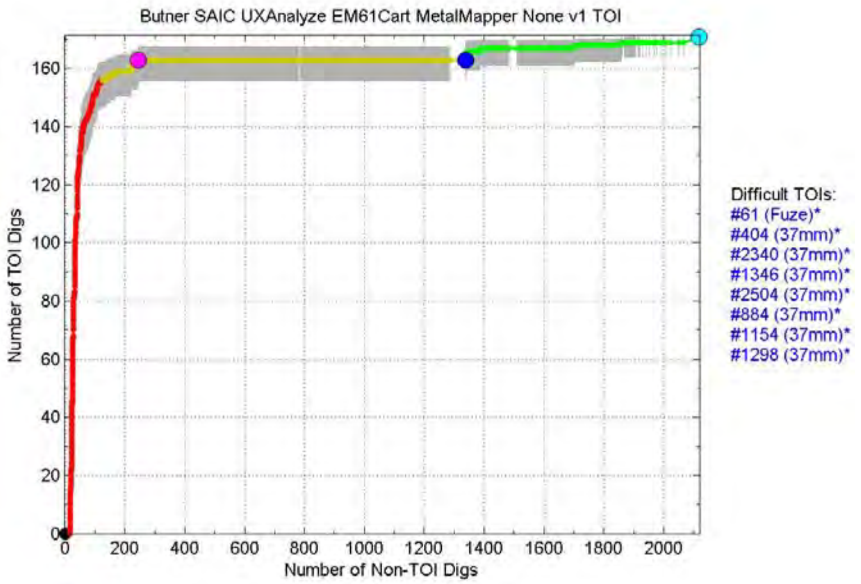


Figure 100: SAIC's performance based on amplitudes, ratios, and decay rates of polarizabilities inverted from the EM61-Mk2 dynamic data and requested MetalMapper static data using the UX-Analyze/IDL multisource dipole model and classified with a rules-based library matcher optimized over an existing library.

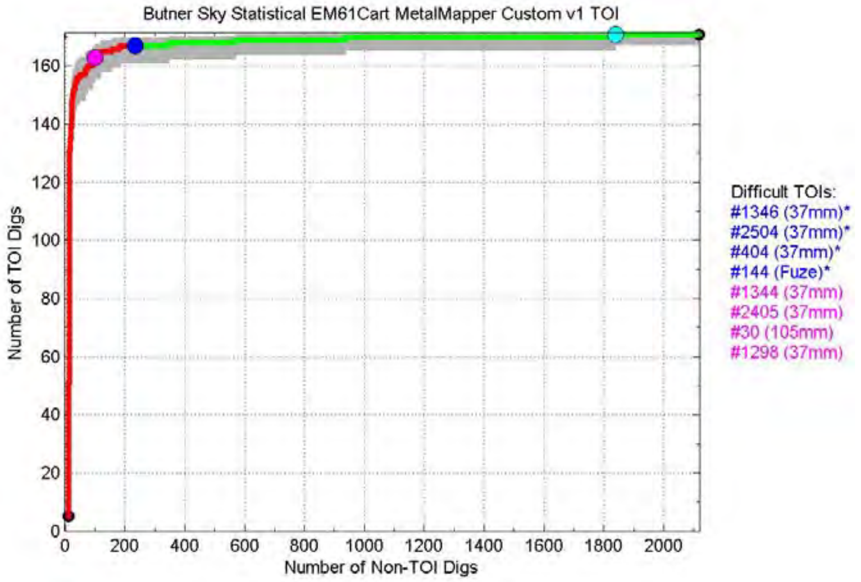


Figure 101: Sky's performance based on polarizabilities inverted from the EM61-Mk2 dynamic data and the requested MetalMapper static data using the UXOLab single-source and two-source dipole models and classified with a statistical classifier optimized over a custom training set.

TEMTADS Static Data

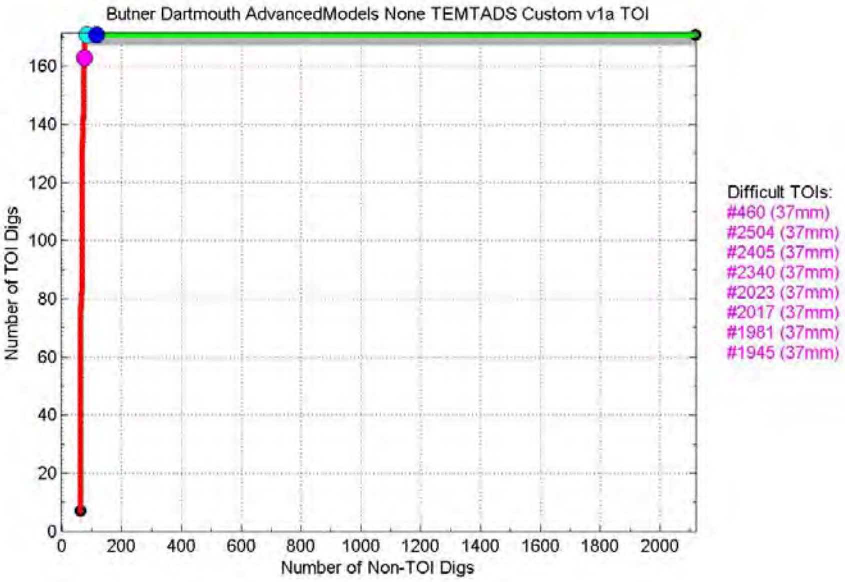


Figure 102: Dartmouth's performance based on full curves inverted from the TEMTADS static data using the OVNMS non-dipole model and classified with a statistical classifier optimized over a custom training set. Identical to Figure 58 in the text.

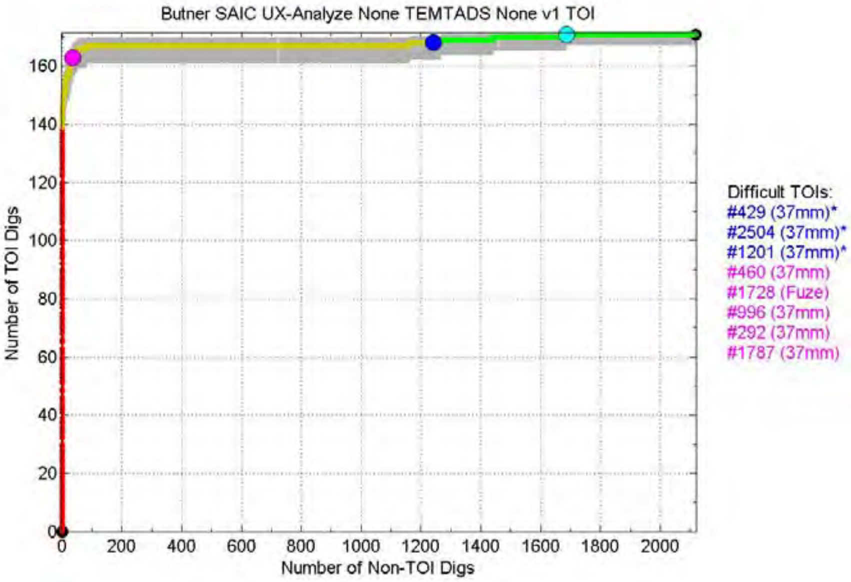


Figure 103: SAIC's performance based on amplitudes, ratios, and decay rates of polarizabilities inverted from the MetalMapper static data using the UX-Analyze/IDL multisource dipole model and classified with a rules-based library matcher optimized over an existing library. Identical to Figure 48 in the text.

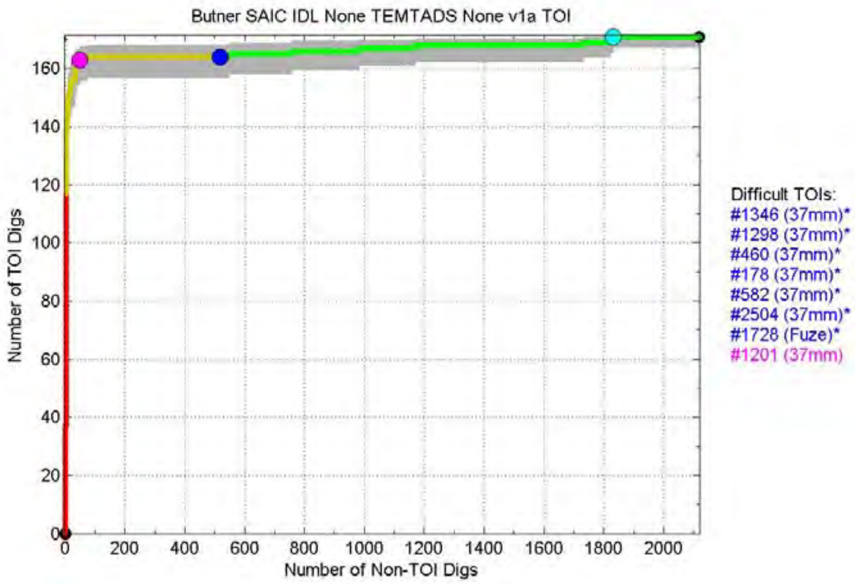


Figure 104: SAIC's performance based on amplitudes, ratios, and decay rates of polarizabilities inverted from the MetalMapper static data using an IDL single-source dipole model and classified with a rules-based library matcher optimized over an existing library. Identical to Figure 57 in the text.

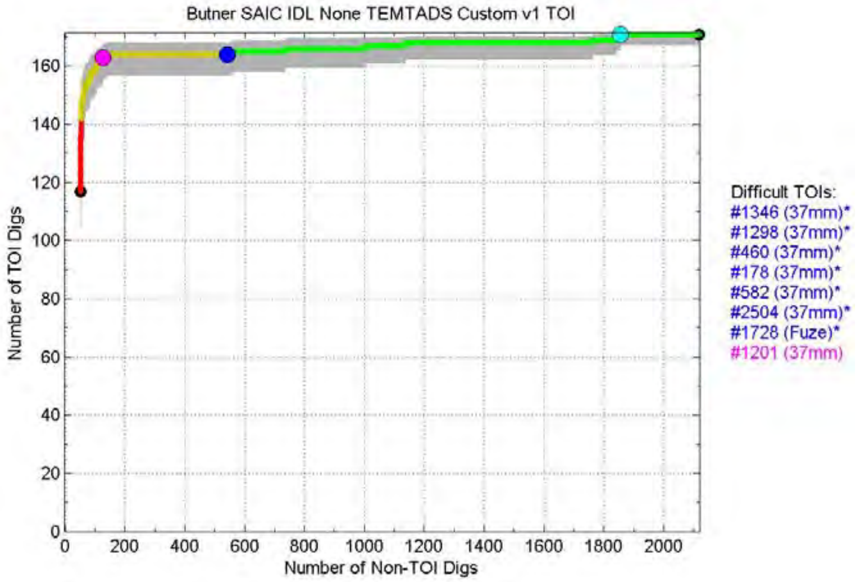


Figure 105: SAIC's performance based on amplitudes, ratios, and decay rates of polarizabilities inverted from the MetalMapper static data using an IDL single-source dipole model and classified with a rules-based library matcher optimized over a custom training set. This second-pass custom training set was built from the "Likely TOI" anomalies in the previous figure.

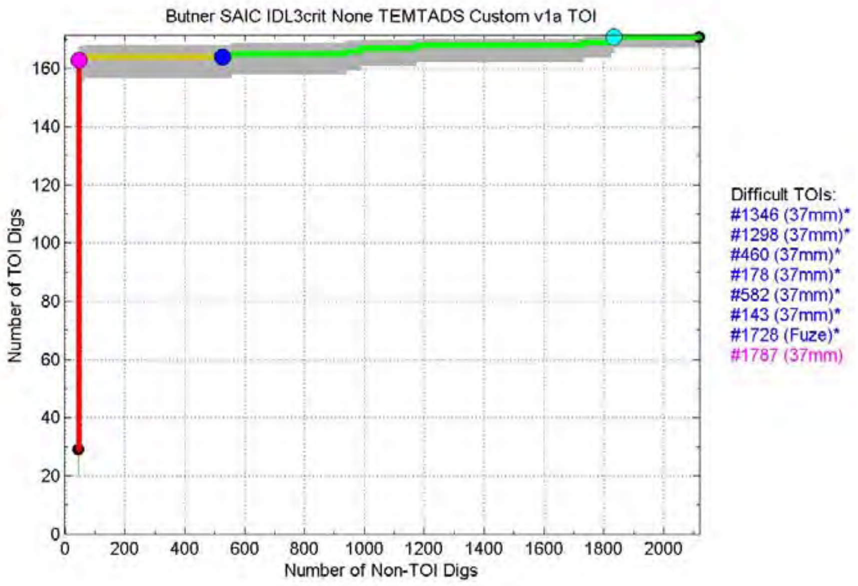


Figure 106: SAIC's performance based on amplitudes, ratios, and decay rates of polarizabilities inverted from the MetalMapper static data using an IDL single-source dipole model and classified with a rules-based library matcher optimized over a custom training set. Two polarizability ratios were used.

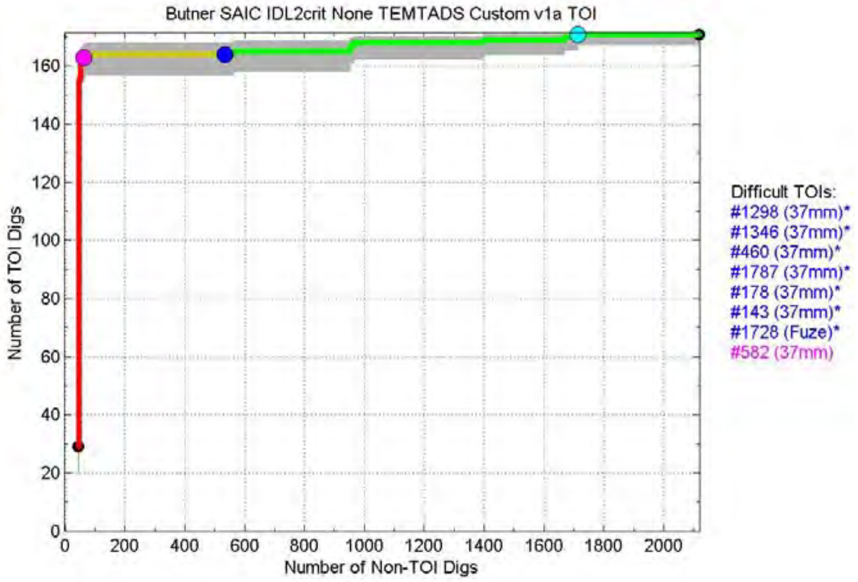


Figure 107: SAIC's performance based on amplitudes, ratios, and decay rates of polarizabilities inverted from the MetalMapper static data using an IDL single-source dipole model and classified with a rules-based library matcher optimized over a custom training set. One polarizability ratio was used.

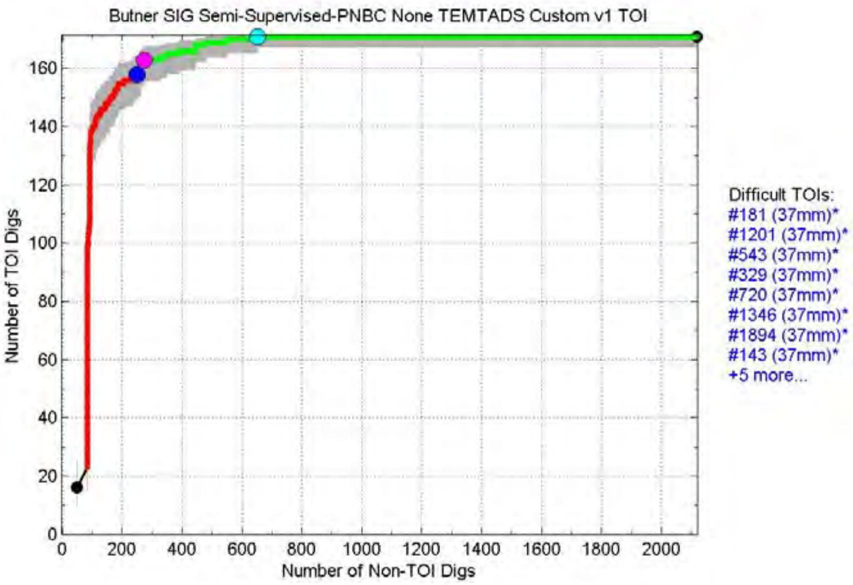


Figure 108: SIG's performance based on its first retrospective analysis of polarizabilities inverted from a dipole model and classified with a statistical classifier optimized over a custom training set.

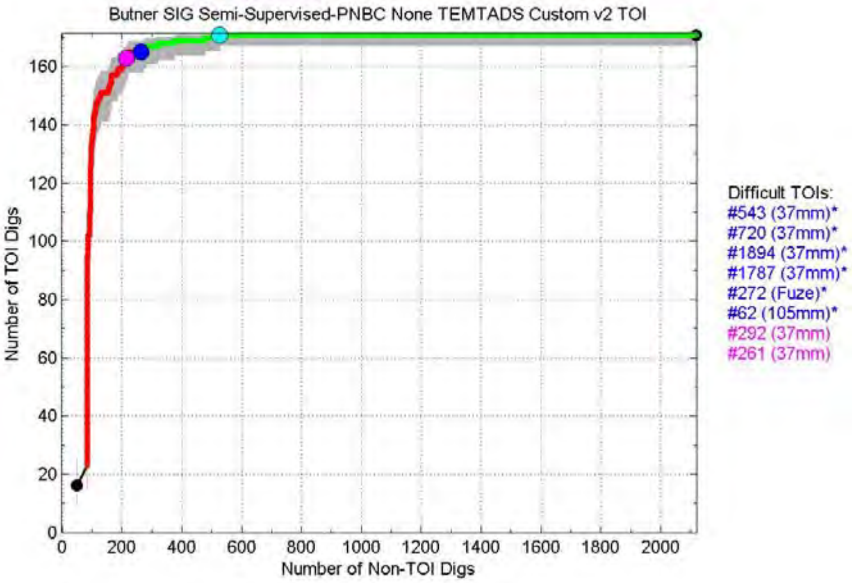


Figure 109: SIG's performance based on its second retrospective analysis of polarizabilities inverted from a dipole model and classified with a statistical classifier optimized over a custom training set.

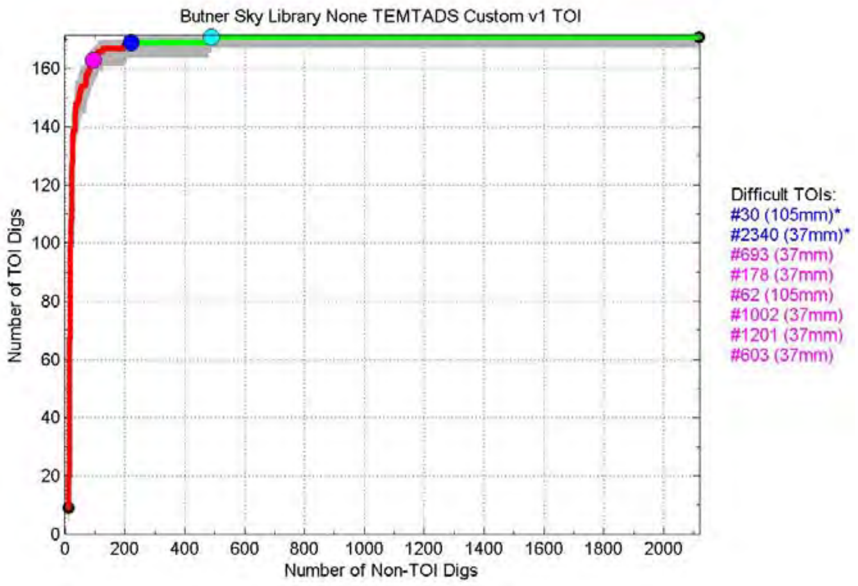


Figure 110: Sky's performance based on full polarizability curves inverted from the TEMTADS static data using the UXOLab single-source and two-source dipole models and classified with a library matcher optimized over a custom training set.

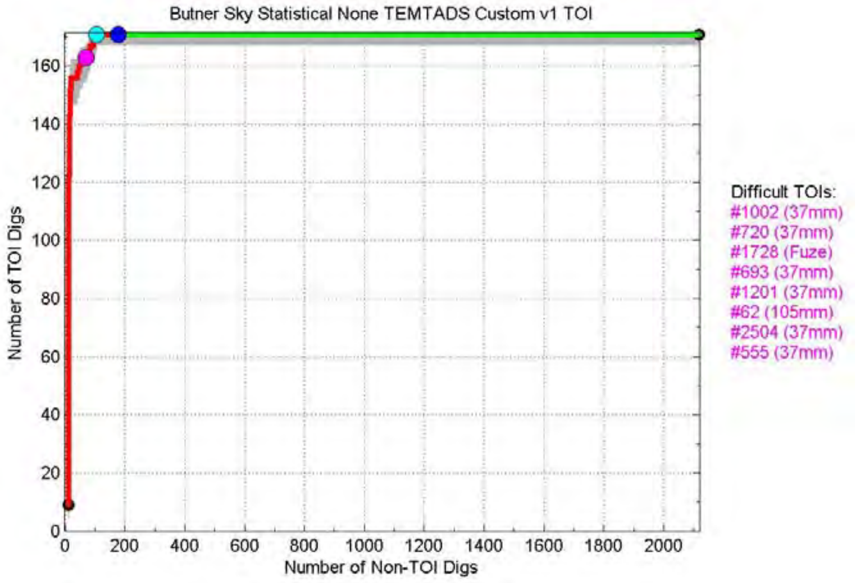


Figure 111: Sky's performance based on full polarizability curves inverted from the TEMTADS static data using the UXOLab single-source and two-source dipole models and classified with a statistical classifier optimized over a custom training set. Identical to Figure 45 in the text.

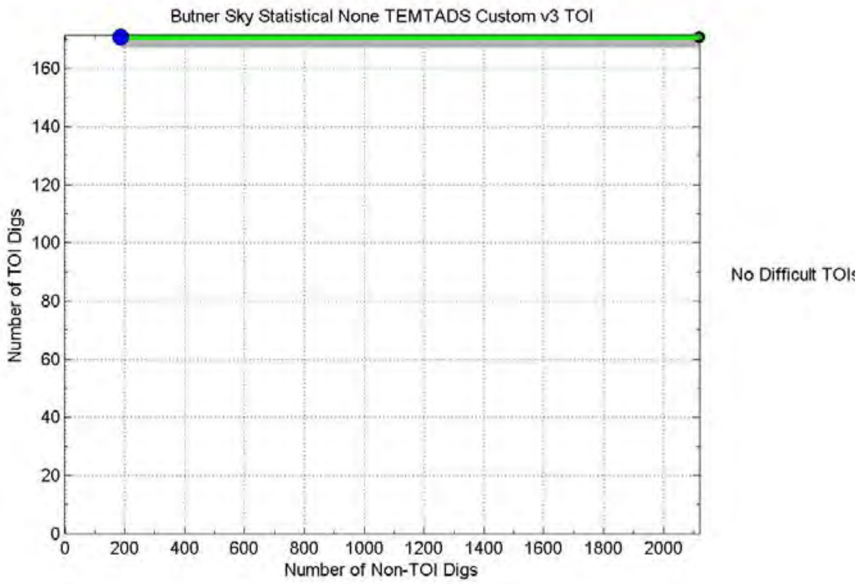


Figure 112: Sky's performance based on full polarizability curves inverted from the TEMTADS static data using the UXOLab single-source and two-source dipole models and classified with a statistical classifier optimized over a custom training set. This second-pass custom training set consisted of those anomalies that fell below the don't dig threshold in the previous figure, as well as (in error) some anomalies that rose above threshold.

EM61-Mk2 Dynamic Data with TEMTADS Static Data Requests

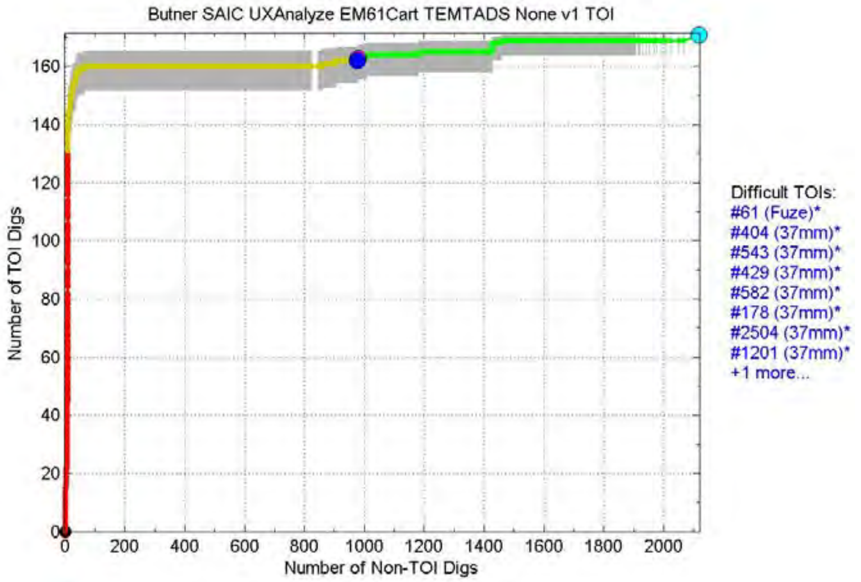


Figure 113: SAIC's performance based on amplitudes, ratios, and decay rates of polarizabilities inverted from the EM61-Mk2 dynamic data and requested TEMTADS static data using the UX-Analyze/IDL multisource dipole model and classified with a rules-based library matcher optimized over an existing library. Identical to Figure 47 in the text.

Appendix B. Classification Metrics

Table 7: Summary statistics for the ranked anomaly lists, including the number of TOI digs (TP), the percentage of TOIs dug (Pd), the number of Non-TOI digs (FP) and the percentage reduction in Non-TOI digs (1-Pfa) at the analyst's don't dig threshold (dark blue dot on classification performance curve). The ranked anomaly lists are ordered first by descending TP and then by ascending FP. There were a total of 171 TOI anomalies and 2119 Non-TOI anomalies. Retrospective analyses are shaded in gray.

Order	Ranked Anomaly List						Results			
	Analyst	Method	Dynamic Data	Static Data	Training Set	Version	Number of TOI Digs (TP)	Percentage of TOIs Dug (Pd)	Number of Non-TOI Digs (FP)	Percentage Reduction in Non-TOI Digs (1-Pfa)
1	Dartmouth	Advanced Models	None	TEMTADS	Custom	1a	171	100%	116	95%
2	Dartmouth	Advanced Models	None	MetalMapper	Custom	2	171	100%	174	92%
3	Sky	Statistical	None	TEMTADS	Custom	1	171	100%	179	92%
4	Sky	Statistical	None	TEMTADS	Custom	3	171	100%	186	91%
5	NAEVA	3crit2crit	None	MetalMapper	Standard	1	171	100%	1005	53%
6	NAEVA	3crit1crit	None	MetalMapper	Standard	1	171	100%	1016	52%
7	SAIC	IDL3crit	None	MetalMapper	Custom	1a	171	100%	1824	14%
8	Parsons	peak4	EM61Cart	None	Custom	3	171	100%	1866	12%
9	SAIC	IDL2crit	None	MetalMapper	Custom	1a	171	100%	1882	11%

Order	Ranked Anomaly List						Results			
	Analyst	Method	Dynamic Data	Static Data	Training Set	Version	Number of TOI Digs (TP)	Percentage of TOIs Dug (Pd)	Number of Non-TOI Digs (FP)	Percentage Reduction in Non-TOI Digs (1-Pfa)
10	NAEVA	Amplitude	EM61Cart	None	Standard	1b	171	100%	2120	0%
11	NAEVA	3crit	None	MetalMapper	Standard	1	170	99%	777	63%
12	Parsons	UXA1	EM61Cart	None	Custom	3	170	99%	1763	17%
13	SAIC	Uxanalyze Rules	EM61Cart	None	Standard	2	170	99%	1801	15%
14	Parsons	peak1	EM61Cart	None	Custom	3	170	99%	1830	14%
15	Sky	Library	None	TEMTADS	Custom	1	169	99%	222	90%
16	Sky	Statistical	None	MetalMapper	Custom	3	169	99%	259	88%
17	Sky	Statistical	MetalMapper	None	Custom	3	169	99%	1152	46%
18	Parsons	peak2	EM61Cart	None	Custom	3	169	99%	1458	31%
19	Sky	Total Polarizability decay and magnitude	EM61Cart	None	None	1a	169	99%	1610	24%
20	SAIC	UX-Analyze	None	TEMTADS	None	1	168	98%	1242	41%
21	Sky	Statistical	None	MetalMapper	Custom	1	167	98%	137	94%
22	Sky	Statistical	EM61Cart	MetalMapper	Custom	1	167	98%	235	89%
23	Sky	Library	None	MetalMapper	Custom	1	167	98%	315	85%
24	CH2MHill	LibMatch-and-HumanInLoop	None	MetalMapper	Custom	1	166	97%	89	96%

Order	Ranked Anomaly List						Results			
	Analyst	Method	Dynamic Data	Static Data	Training Set	Version	Number of TOI Digs (TP)	Percentage of TOIs Dug (Pd)	Number of Non-TOI Digs (FP)	Percentage Reduction in Non-TOI Digs (1-Pfa)
25	Parsons	UXAnalyze3	EM61Cart	MetalMapper	Custom	3	166	97%	679	68%
26	Parsons	UXAnalyze4	EM61Cart	MetalMapper	Custom	1	166	97%	694	67%
27	SIG	Semi-Supervised-PNBC	None	TEMTADS	Custom	2	165	96%	264	88%
28	Geometrics	Rules+ANN+LM	None	MetalMapper	Standard	1b	165	96%	327	85%
29	CH2MHill	SimpleLibMatch	None	MetalMapper	Custom	1	165	96%	337	84%
30	NAEVA	Tau123RuleBased	EM61Cart	None	Standard	1b	165	96%	1521	28%
31	SAIC	IDL	None	TEMTADS	None	1a	164	96%	519	76%
32	SAIC	IDL3crit	None	TEMTADS	Custom	1a	164	96%	526	75%
33	SAIC	IDL2crit	None	TEMTADS	Custom	1a	164	96%	533	75%
34	SAIC	IDL	None	TEMTADS	Custom	1	164	96%	543	74%
35	SAIC	Rules	None	MetalMapper	Custom	1	164	96%	1009	52%
36	Parsons	UXAnalyze3	EM61Cart	MetalMapper	Custom	2	163	95%	603	72%
37	Parsons	UXAnalyze1	EM61Cart	MetalMapper	Custom	2	163	95%	1019	52%
38	SAIC	UXAnalyze	EM61Cart	MetalMapper	None	1	163	95%	1337	37%
39	SAIC	UXAnalyze	EM61Cart	TEMTADS	None	1	162	95%	977	54%

Order	Ranked Anomaly List						Results			
	Analyst	Method	Dynamic Data	Static Data	Training Set	Version	Number of TOI Digs (TP)	Percentage of TOIs Dug (Pd)	Number of Non-TOI Digs (FP)	Percentage Reduction in Non-TOI Digs (1-Pfa)
40	Parsons	UXAnalyze2	EM61Cart	MetalMapper	Custom	2	161	94%	828	61%
41	Parsons	UXA2	EM61Cart	None	Custom	3	160	94%	1036	51%
42	Parsons	peak3	EM61Cart	None	Custom	3	160	94%	1187	44%
43	Parsons	peak5	EM61Cart	None	Custom	3	160	94%	1194	44%
44	Parsons	UXA4	EM61Cart	None	Custom	3	159	93%	1068	50%
45	SIG	Semi-Supervised-PNBC	None	TEMTADS	Custom	1	158	92%	251	88%
46	Sky	Man-in-Loop	MetalMapper	None	Custom	1	158	92%	740	65%
47	SAIC	Rules	None	MetalMapper	None	1a	158	92%	998	53%
48	Sky	Library	MetalMapper	None	Custom	1	156	91%	524	75%
49	Sky	Statistical	MetalMapper	None	Custom	1	152	89%	573	73%
50	Geometrics	Rules+ANN	None	MetalMapper	Standard	1b	150	88%	238	89%
51	Geometrics	ANN	None	MetalMapper	Standard	1b	150	88%	250	88%
52	NAEVA	Tau234Size RuleBased	EM61Cart	None	Standard	1b	133	78%	701	67%
53	Parsons	UXA5	EM61Cart	None	Custom	3	127	74%	872	59%
54	Parsons	UXA3	EM61Cart	None	Custom	3	127	74%	873	59%

Table 8: Summary statistics for the ranked anomaly lists, including the number of TOI digs (TP), the percent of TOIs dug (Pd), the number of Non-TOI digs (FP) and the percent reduction in Non-TOI digs (1-Pfa) at the retrospective “best” don’t dig threshold (light blue dot on classification performance curve). The ranked anomaly lists have been ordered by ascending FP. The “best” don’t dig threshold was that for which all TOIs were dug with the minimum number of (i.e., greatest reduction in) Non-TOI digs. There were a total of 171 Non-TOI anomalies and 2119 Non-TOI anomalies. Retrospective analyses are shaded in gray.

Order	Ranked Anomaly List						Results			
	Analyst	Method	Dynamic Data	Static Data	Training Set	Version	Number of TOI Digs (TP)	Percent of TOIs Dug (Pd)	Number of Non-TOI Digs (FP)	Percent Reduction in Non-TOI Digs (1-Pfa)
1	Dartmouth	AdvancedModels	None	TEMTADS	Custom	1a	171	100%	85	96%
2	Sky	Statistical	None	TEMTADS	Custom	1	171	100%	106	95%
3	Dartmouth	AdvancedModels	None	MetalMapper	Custom	2	171	100%	149	93%
4	Sky	Statistical	None	TEMTADS	Custom	3	171	100%	186	91%
5	Sky	Library	None	TEMTADS	Custom	1	171	100%	488	77%
6	SIG	Semi-Supervised-PNBC	None	TEMTADS	Custom	2	171	100%	525	75%
7	SIG	Semi-Supervised-PNBC	None	TEMTADS	Custom	1	171	100%	653	69%
8	NAEVA	3crit1crit	None	MetalMapper	Standard	1	171	100%	807	62%
9	NAEVA	3crit2crit	None	MetalMapper	Standard	1	171	100%	868	59%
10	CH2MHill	SimpleLibMatch	None	MetalMapper	Custom	1	171	100%	985	54%
11	NAEVA	3crit	None	MetalMapper	Standard	1	171	100%	1028	52%
12	CH2MHill	LibMatch-and-HumanInLoop	None	MetalMapper	Custom	1	171	100%	1250	41%

Order	Ranked Anomaly List						Results			
	Analyst	Method	Dynamic Data	Static Data	Training Set	Version	Number of TOI Digs (TP)	Percent of TOIs Dug (Pd)	Number of Non-TOI Digs (FP)	Percent Reduction in Non-TOI Digs (1-Pfa)
13	Sky	Library	MetalMapper	None	Custom	1	171	100%	1266	40%
14	SAIC	Rules	None	MetalMapper	Custom	1	171	100%	1334	37%
15	NAEVA	Amplitude	EM61Cart	None	Standard	1b	171	100%	1364	36%
16	Sky	Statistical	MetalMapper	None	Custom	3	171	100%	1403	34%
17	Sky	Statistical	MetalMapper	None	Custom	1	171	100%	1403	34%
18	Sky	Man-in-Loop	MetalMapper	None	Custom	1	171	100%	1543	27%
19	NAEVA	Tau234Size RuleBased	EM61Cart	None	Standard	1b	171	100%	1644	22%
20	Sky	Statistical	None	MetalMapper	Custom	1	171	100%	1668	21%
21	Sky	Statistical	None	MetalMapper	Custom	3	171	100%	1669	21%
22	SAIC	Rules	None	MetalMapper	None	1a	171	100%	1673	21%
23	SAIC	UX-Analyze	None	TEMTADS	None	1	171	100%	1687	20%
24	SAIC	IDL2crit	None	TEMTADS	Custom	1a	171	100%	1712	19%
25	SAIC	IDL3crit	None	MetalMapper	Custom	1a	171	100%	1787	16%
26	Parsons	peak4	EM61Cart	None	Custom	3	171	100%	1789	16%
27	NAEVA	Tau123RuleBased	EM61Cart	None	Standard	1b	171	100%	1817	14%

Order	Ranked Anomaly List						Results			
	Analyst	Method	Dynamic Data	Static Data	Training Set	Version	Number of TOI Digs (TP)	Percent of TOIs Dug (Pd)	Number of Non-TOI Digs (FP)	Percent Reduction in Non-TOI Digs (1-Pfa)
28	SAIC	IDL2crit	None	MetalMapper	Custom	1a	171	100%	1818	14%
29	SAIC	IDL	None	TEMTADS	None	1a	171	100%	1832	14%
30	SAIC	IDL3crit	None	TEMTADS	Custom	1a	171	100%	1834	13%
31	Sky	Statistical	EM61Cart	MetalMapper	Custom	1	171	100%	1839	13%
32	SAIC	IDL	None	TEMTADS	Custom	1	171	100%	1856	12%
33	Parsons	UXAnalyze4	EM61Cart	MetalMapper	Custom	1	171	100%	1865	12%
34	Parsons	UXA1	EM61Cart	None	Custom	3	171	100%	1873	12%
35	Parsons	UXA2	EM61Cart	None	Custom	3	171	100%	1873	12%
36	Parsons	UXA5	EM61Cart	None	Custom	3	171	100%	1873	12%
37	Parsons	UXA3	EM61Cart	None	Custom	3	171	100%	1873	12%
38	Parsons	peak1	EM61Cart	None	Custom	3	171	100%	1939	9%
39	Parsons	peak2	EM61Cart	None	Custom	3	171	100%	1939	9%
40	Parsons	peak3	EM61Cart	None	Custom	3	171	100%	1939	9%
41	Parsons	peak5	EM61Cart	None	Custom	3	171	100%	1943	8%
42	Sky	total polarizability decay and magnitude	EM61Cart	None	None	1a	171	100%	1944	8%

Order	Ranked Anomaly List						Results			
	Analyst	Method	Dynamic Data	Static Data	Training Set	Version	Number of TOI Digs (TP)	Percent of TOIs Dug (Pd)	Number of Non-TOI Digs (FP)	Percent Reduction in Non-TOI Digs (1-Pfa)
43	Parsons	UXAnalyze2	EM61Cart	MetalMapper	Custom	2	171	100%	1961	8%
44	Parsons	UXA4	EM61Cart	None	Custom	3	171	100%	1975	7%
45	Parsons	UXAnalyze1	EM61Cart	MetalMapper	Custom	2	171	100%	1984	6%
46	Parsons	UXAnalyze3	EM61Cart	MetalMapper	Custom	3	171	100%	1994	6%
47	SAIC	UxanalyzeRules	EM61Cart	None	Standard	2	171	100%	2001	6%
48	Sky	Library	None	MetalMapper	Custom	1	171	100%	2106	1%
49	Parsons	UXAnalyze3	EM61Cart	MetalMapper	Custom	2	171	100%	2113	0%
50	SAIC	UXAnalyze	EM61Cart	MetalMapper	None	1	171	100%	2119	0%
51	SAIC	UXAnalyze	EM61Cart	TEMTADS	None	1	171	100%	2119	0%
52	Geometrics	Rules+ANN+LM	None	MetalMapper	Standard	1b	171	100%	2120	0%
53	Geometrics	Rules+ANN	None	MetalMapper	Standard	1b	171	100%	2120	0%
54	Geometrics	ANN	None	MetalMapper	Standard	1b	171	100%	2120	0%

Figures

Tables

References

[1] —. *2010 ESTCP Classification Study, Camp Butner, NC: Former Camp Butner: SAIC Data Analysis Demonstration Plan*. SAIC. 10 August 2010.

[2] —. *2010 ESTCP Classification Study: Former Camp Butner*. ESTCP. 28 May 2010.

[3] —. *2010 ESTCP UXO Classification Study, Rougemont, NC: Draft Demonstration Data Report: Former Camp Butner: MTADS Discrimination Array (TEMTADS) Survey*. Nova Geophysics Inc. 6 October 2010.

[4] —. *Advisory Group Update on Butner Results*. ESTCP. 20 September 2010.

[5] —. *Amplitude EM61 Cart Prioritization - Decision Memo.* NAEVA Geophysics Inc. 15 November 2010.

[6] —. *Classification Memo for the MetalMapper Advanced Sensor: Three Criteria Considered for All Targets (3crit).* NAEVA Geophysics Inc. 27 January 2011.

[7] —. *Data Collection Report: ESTCP UXO Classification Study: EM61-Mk2 Data Collection and Analysis at Camp Butner, Durham, NC.* HydroGeoLogic Inc. May 2010.

[8] —. *Data Collection Report: ESTCP UXO Discrimination Study Site Evaluation: EM61-Mk2 Data Collection and Analysis at Camp Butner, Durham, NC.* HydroGeoLogic Inc. October 2009.

[9] —. *Decay Constant-Based Ranked Dig Lists.* Parsons Inc. 3 February 2011.

[10] —. *Decision Memo.* SAIC. 15 October 2010.

[11] —. *Decision Memo.* SAIC. 18 October 2010.

[12] —. *Decision Memo.* SAIC. 22 October 2010.

[13] —. *Decision Memo.* SAIC. 11 November 2010a.

[14] —. *Decision Memo.* SAIC. 11 November 2010b.

[15] —. *Decision Memo for the MetalMapper Advanced Sensor: Two Stage Classification using Three or One Criteria (3crit1crit).* NAEVA Geophysics Inc. 27 January 2011.

[16] —. *Decision Memo for the MetalMapper Advanced Sensor: Two Stage Classification using Three or Two Criteria (3crit2crit).* NAEVA Geophysics Inc. 27 January 2011.

[17] —. *Metal Detectors.* Geonics Limited. http://www.geonics.com/pdfs/downloads/metaldetectors.pdf. Accessed 11 July 2011.

[18] —. *EM61-Mk2 Cart Data Analysis: Former Camp Butner*. NAEVA Geophysics Inc. 13 August 2010.

[19] —. *Final Engineering Evaluation/Cost Analysis, Former Camp Butner, Camp Butner, NC.* Parsons Inc. July 2004.

[20] —. *Former Camp Butner, North Carolina: Munitions and Explosives of Concern Surface Clearance Site Report.* HydroGeoLogic Inc. 4 May 2010.

[21] —. *Intrusive Investigation Report: ESTCP UXO Discrimination Study Site Evaluation: EM61-MK2 Data Collection and Analysis at Camp Butner, Durham, NC.* HydroGeoLogic Inc. October 2009.

[22] —. *Planned Approach For ESTCP Camp Butner Discrimination Study.* Sky Research Inc. 2 August 2010.

[23] —. *Program Areas: Munitions Response: Land.* SERDP/ESTCP. http://www.serdp.org/Program-Areas/Munitions-Response/Land. Accessed 8 July 2011.

[24] —. *Report of the Defense Science Board Task Force on Unexploded Ordnance.* Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics. 2003.

[25] —. *Tau123 Rule Based EM61 Cart Prioritization – Decision Memo.* NAEVA Geophysics Inc. 15 November 2010.

[26] —. *Tau234 Size Rule Based EM61 Cart Prioritization – Decision Memo.* NAEVA Geophysics Inc. 15 November 2010.

[27] —. *Unexploded Ordnance (UXO).* U.S. Army Environmental Command. http://aec.army.mil/usaec/technology/uxo00.html. Accessed 8 July 2011.

[28] —. *White Paper: Planned Approach for ESTCP Camp Butner Discrimination Study.* Sky Research Inc. 6 August 2010.

[29] Andrews, A., et al. *ESTCP Pilot Program: Classification Approaches in Munitions Response: Camp Butner, North Carolina.* ESTCP. June 2011.

[30] Bishop, C. M. *Neural Networks for Pattern Recognition.* Oxford University Press. 1999.

[31] Cazares, S. *Analysis Plan for the UXO Classification Study at the Former Camp Butner.* #NS D-4182. Institute for Defense Analyses. 27 August 2010.

[32] Cazares, S. *Seed Plan for the UXO Classification Study at the Former Camp Butner.* #NS D-4125. Institute for Defense Analyses. 19 July 2010.

[33] Cazares, S., et al. *The UXO Classification Study at the Former Camp Sibert.* #D-3572. Institute for Defense Analyses. 1 January 2009.

[34] Cazares, S. and Tuley, M. *The UXO Classification Demonstration at San Luis Obispo, CA.* #D-4148. Institute for Defense Analyses. 1 September 2010.

[35] Dasgupta, N. Personal Communication. 16 February 2011.

[36] Keiswetter, D. “Classification with EM61 Data.” *Partners in Environmental Technology Technical Workshop and Symposium, Short Course 1: Advances in Classification Methods for Military Munitions Response.* 1 December 2001.

[37] Keiswetter, D. Personal Communication. 6 July 2011.

[38] Keiswetter, D. “SAIC Analysis of Data Acquired at Camp Butner, NC.” *Partners in Environmental Technology Technical Workshop and Symposium, Technical Session 2D: Classification Methods for Military Munitions Response.* 1 December 2010.

[39] Keiswetter, D. “Technical Transfer Demonstration: Camp Butner Analysis.” *ESTCP Munitions Management In-Progress Review.* 9 February 2011.

[40] Khadr, N. Personal Communication. 10 May 2010.

[41] Khadr, N. Personal Communication. 13 May 2010.

[42] Klaff, T. “Discrimination Applied to EMI Data Collected at the Former Camp Butner.” *ESTCP Munitions Management In-Progress Review*. 9 February 2011.

[43] Macshassy S. and Provost F. “Confidence Bands for ROC Curves: Methods and Empirical Study.” *Proc 1st Workshop on ROC Analysis in AI*. 2004.

[44] Murray, C. Personal Communication. 28 January 2011.

[45] Murray, C. Personal Communication. 2 February 2011.

[46] Murray, C. “Technology Transfer Demonstration: Former Camp Butner.” *ESTCP Munitions Management In-Progress Review*. 9 February 2011.

[47] Paski, A. “Camp Butner Data Collection and Analysis.” *ESTCP Munitions Management In-Progress Review*. 9 February 2011.

[48] Paski, A., et al. “Former Camp Butner Site Description and EM61 Data Collection and Analysis.” *Partners in Environmental Technology Technical Symposium and Workshop, Technical Session 2D: Classification Methods for Military Munitions Response*. 1 December 2010.

[49] Pasion, L. Personal Communication. 15 June 2011.

[50] Pasion, L. “Practical Strategies for UXO Discrimination: Camp Butner Analysis.” *ESTCP Munitions Management In-Progress Review*. 9 February 2011.

[51] Pasion, L., et al. “UXO Discrimination Using Full Coverage and Cued Interrogation Data Sets at Camp Butner, NC.” *Partners in Environmental Technology Technical Workshop and Symposium, Technical Session 2D: Classification Methods for Military Munitions Response*. 1 December 2010.

[52] Prouty, M. *Data Analysis Plan: Former Camp Butner*. Geometrics Inc. August 2010.

[53] Prouty, M. *Data Collection Report: MetalMapper System: Camp Butner Discrimination Study*. Geometrics Inc. 27 August 2010.

[54] Prouty, M. “METALMAPPER: A Multi-Sensor TEM and Magnetic Gradiometer System for UXO Detection and Classification.” *ESTCP Munitions Management In-Progress Review*. 9 February 2011.

[55] Prouty, M. *Training Memo: Detection and Classification with the MetalMapper at the Former Camp Butner, NC*. Geometrics Inc. 14 September 2010.

[56] Shubitidze, F. “A Complex Approach to UXO Discrimination: Combining Advanced EMI Forward Models and Statistical Signal Processing Methodologies.” *ESTCP Munitions Management In-Progress Review*. 9 February 2011.

[57] Shubitidze, F. *Advanced EMI Models for ESTCP Live UXO Site Classification Studies: Camp Butner Metal Mapper Data Sets*. Sky Research Inc. November 2010.

[58] Shubitidze, F. *Advanced EMI Models for ESTCP Live UXO Site Classification Studies: Camp Butner TEMTADS Data Sets*. Sky Research Inc. November 2010.

[59] Shubitidze, F. “Camp Butner UXO Data Inversion and Classification Using Advanced EMI Models.” *Partners in Environmental Technology Technical Workshop and Symposium, Technical Session 2D: Classification Methods for Military Munitions Response*. 1 December 2010.

[60] Snyder, S., et al. “UXO Detection & Classification with the MetalMapper at Camp Butner, NC.” *Partners in Environmental Technology Technical Workshop*

and Symposium, Technical Session 2D: Classification Methods for Military Munitions Response. 1 December 2010.

[61] Steinhurst, D. “UXO Classification Using EMI Sensors: UXO Classification Study: Former Camp Butner, NC.” *ESTCP Munitions Management In-Progress Review*. 9 February 2011.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE
July 2011

2. REPORT TYPE
Final

3. DATES COVERED *(From–To)*
May 2010 – April 2011

4. TITLE AND SUBTITLE

The UXO Classification Demonstration
at the Former Camp Butner, NC

5a. CONTRACT NUMBER
DASW01 04 C 0003

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Shelley Cazares
Michael Tuley
Elizabeth Ayers

5d. PROJECT NUMBER

5e. TASK NUMBER
AM-2-1528

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER

IDA Document D-4379
Log: H11-001211/1

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Environmental Security Technology Certification Program
901 N. Stuart Street, Suite 303
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The Environmental Security Technology Certification Program (ESTCP) carried out the third live-site UXO classification demonstration at the former Camp Butner, NC in 2010. The main goal of the demonstration was to test and validate currently available and emerging classification technologies on a live site under operational conditions. Another goal was to involve environmental regulators, program managers, and other stakeholders in the design, execution, and evaluation of the demonstration to better understand what might be required in a real world remediation project if detected targets were classified as clutter and therefore left in the ground. This report provides a detailed record of the scoring of the detection and discrimination performance of all the demonstrators, sensors and algorithms involved in the demonstration.

15. SUBJECT TERMS

Unexploded Ordnance (UXO), discrimination, detection, classification, electromagnetic induction sensors

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Dr. Herbert Nelson
a. REPORT Uncl.	b. ABSTRACT Uncl.	c. THIS PAGE Uncl.	SAR	156	19b. TELEPHONE NUMBER *(include area code)* 703-696-8726